How Multi-Head Transformers Specialize for In-Context Learning: A Theoretical Analysis of Training Dynamics and Generalization

Anonymous authors
Paper under double-blind review

Abstract

In-context learning (ICL) enables transformers to infer task-specific mappings from demonstrations in the prompt. A theoretical understanding of how multiple attention heads learn to specialize during training remains limited. We study the training dynamics and generalization of a multi-head softmax transformer on compositional ICL tasks, where each input consists of multiple latent components that are independently mapped to corresponding output components. We jointly analyze the training of attention and value parameters, and show that their coupled gradient-based dynamics amplify weak initial differences between heads into complementary specialization: each head learns to match and retrieve information according to a different latent component. We establish finite-step convergence and show that the learned model generalizes to unseen combinations of component mappings and to out-of-domain pattern families. Our theoretical insights are validated on synthetic and real experiments.

1 Introduction

In-context learning (ICL) enables Transformers to infer a task from demonstrations provided in the prompt and make predictions on new queries without updating their parameters. This capability allows a single pretrained model to adapt to a wide range of tasks through the context alone, and has become one of the defining capabilities of modern large language models. The empirical success of ICL raises a fundamental question of how Transformers acquire and implement such context-dependent learning behavior (Garg et al., 2022).

A line of theoretical work interprets ICL through the learning algorithms implemented by Transformer forward passes. Transformers and related attention models have been shown to implement gradient-based learning algorithms and their variants (Akyürek et al., 2023; von Oswald et al., 2023; Ahn et al., 2023; Mahankali et al., 2024). Bai et al. (2023) establishes broader classes of statistical learning and algorithm-selection capabilities. Another line of works studies how ICL capabilities are acquired through training, establishing optimization dynamics and generalization for trained attention models (Zhang et al., 2024; Huang et al., 2024).

Multi-head attention is a central architectural component of modern Transformers, yet its role in ICL remains much less understood. Multi-head attention is known to improve in-context risk (Cui et al., 2025), but not how this structure arises through training. Existing theory has largely either characterized the learning behavior of the Transformer as a whole or analyzed how ICL emerges through training in single-head models. Neither perspective explains how multiple heads interact and differentiate during training. We therefore ask: *How does gradient-based training organize multiple attention heads into specialized roles that support ICL generalization?*

A few recent works investigate the training and specialization of multi-head attention. Chen et al. (2024a;b) establish forms of head specialization through task allocation or specialized roles in structured sequence models. He et al. (2025) characterize joint query–key and output–value dynamics and show how multiple heads collectively implement a debiased regression rule. Zhang et al. (2025) characterize progressive learning dynamics in multi-head linear attention. However, these works do not establish how complementary specialization across heads supports generalization to new ICL tasks. Yang et al. (2024) establish convergence and generalization of trained multi-head Trans-

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.

formers to unseen tasks, but do not characterize explicit specialization across heads. Thus, how training-induced specialization across heads supports systematic ICL generalization remains open.

1.1 Our Contributions

This paper addresses this question through a compositional ICL model in which each task consists of multiple component-wise input–output mappings. We analyze the coupled training dynamics of the attention and value parameters of a two-head softmax Transformer, showing how training induces complementary specialization across heads and how this specialization supports ICL generalization. We focus on two heads to isolate the specialization mechanism, while the analysis extends naturally to more heads. The distinctions of this paper from the most relevant works are summarized in Table 1. Our main contributions are as follows.

- **Training dynamics and complementary head specialization.** We provide a finite-step analysis of the coupled gradient-based training dynamics of the attention and value parameters in a two-head softmax Transformer. We show that training amplifies weak initial differences between heads into complementary specialization: different heads learn to match context examples according to different input components and retrieve the corresponding output components. We characterize this specialization analytically through both the learned parameters and the attention patterns.
- **ICL generalization to unseen tasks.** We show that the trained model can solve unseen tasks using only demonstrations provided in the context, even when their complete component mappings have not appeared during training. It suffices that every individual input–output correspondence occurring in the task family is observed in at least one training task. For the full family of up to $(M!)^2$ tasks, only M appropriately chosen training tasks suffice to provide this correspondence-level coverage. Moreover, the required context length scales only linearly in the number of patterns M per component, up to logarithmic factors, rather than the M^2 scaling suggested by jointly matching both components under the prior single-head ICL analysis (Li et al., 2024a).
- **Out-of-domain (OOD) ICL generalization to unseen pattern families.** We further establish OOD ICL generalization to new task families composed of previously unseen input and output patterns obtained through structured transformations of the training pattern families. Thus, the learned ICL mechanism extends beyond unseen mappings over the training patterns to structured shifts in the patterns themselves.

Table 1: Comparison with most relevant theoretical analyses of multi-head ICL.

Work	Softmax	Finite S/GD	Joint QK–OV training	Specialization	Generalization
Chen et al. (2024a)	✓	×	✓	✓	×
Chen et al. (2024b)	✓	×	×	✓	×
He et al. (2025)	✓	×	✓	✓	×
Zhang et al. (2025)	×	×	✓	×	×
Yang et al. (2024)	✓	×	✓	×	Unseen tasks
Ours	✓	✓	✓	✓	Compositional+OOD

Finite S/GD: finite sample stochastic/gradient descent. "Composition" denotes unseen combinations of component mappings, and OOD denotes unseen pattern families.

1.2 Other Related Works

Theoretical studies of in-context learning. Beyond the works discussed above, a broad theoretical literature studies the mechanisms, training, and generalization of ICL. These include analyses of nearest-neighbor and pattern-matching mechanisms (Collins et al., 2024; Li et al., 2024b), feature and representation learning (Kim & Suzuki, 2024), nonlinear ICL under richer task structures (Li et al., 2024a), and recovery of latent causal structure from early gradients(Nichani et al., 2024). These analyses concern single-head attention.

Empirical and mechanistic studies of attention-head specialization. Empirical and mechanistic studies have identified specialized attention heads that support ICL. Early work found that induction heads perform pattern-matching and copying operations that emerge during training (Olsson et al., 2022; Bietti et al., 2023; Reddy, 2024; Edelman et al., 2024). Later studies showed that ICL-related computations need not be localized to a single head, but can be distributed or redundant across

multiple heads and sub-circuits (Singh et al., 2024). Studies of larger pretrained models have further identified heads associated with higher-level ICL functions, including semantic induction and function-vector representations (Ren et al., 2024; Yin & Steinhardt, 2025).

Optimization and generalization of Transformers beyond ICL. A broader theoretical literature studies the optimization and generalization of attention-based models outside the ICL setting. This includes analyses of feature and spatial-structure learning in attention models (Jelassi et al., 2022; Li et al., 2023a), optimization of attention and prompt parameters (Oymak et al., 2023; Tarzanagh et al., 2023), and Transformer training dynamics in broader sequence-modeling settings (Tian et al., 2023a; Li et al., 2023b; Tian et al., 2023b).

2 Problem Formulation

In-context learning (ICL) allows a model to predict the output of a query input x_q using examples provided in the prompt P, without updating the model parameters at the inference time. Specifically, let a task τ, drawn from a task family $\mathcal{T}$, specify an input–output mapping g^τ. A prompt P associated with task τ consists of L context examples followed by a query,

$$P = [\,p_1, \ldots, p_L, p_q\,] \in \mathbb{R}^{d\times(L+1)}, \quad \text{where } p_\ell = \begin{bmatrix} x_\ell \\ y_\ell \end{bmatrix}, \ell \in [L], \quad p_q = \begin{bmatrix} x_q \\ \mathbf{0} \end{bmatrix} \tag{1}$$

Here input-output pair $x_\ell \in \mathbb{R}^{d_x}$, $y_\ell = g^\tau(x_\ell) \in \mathbb{R}^{d_y}$, and $d = d_x + d_y$. The query token p_q contains the query input x_q with zeros in place of the withheld output.

Given a prompt P, a model parameterized by Ψ predicts the query output as

$$\widehat{y}_q = F(\Psi; P). \tag{2}$$

Since the task varies across prompts, the model must learn to infer the task-specific mapping from the context examples and apply it to the query.

Multi-head attention model. We consider F as a one-layer, H-head softmax attention model. For each head h, let W_Q^h, W_K^h, W_V^h, and W_O^h denote its query, key, value, and output matrices, respectively. The model prediction is

$$F(\Psi; P) = \sum_{h=1}^{H} W_O^h W_V^h \sum_{\ell=1}^{L} \mathrm{softmax}\left(\frac{1}{\sqrt{d}}(W_K^h p_\ell)^\top W_Q^h p_q\right) p_\ell, \tag{3}$$

where the softmax is taken over the L context tokens, i.e., $\mathrm{softmax}(z_\ell) = \frac{\exp(z_\ell)}{\sum_{r=1}^{L}\exp(z_r)}$.

To simplify the analysis while preserving the key mechanisms, we directly parameterize each head by $W_{\mathrm{att}}^h := \frac{1}{\sqrt{d}} W_K^{h\top} W_Q^h$, and $W_{\mathrm{OV}}^h := W_O^h W_V^h$, with $\Psi = \{W_{\mathrm{att}}^h, W_{\mathrm{OV}}^h\}_{h\in[H]}$. Here, W_{att}^h determines how head h weights the context examples, while W_{OV}^h determines how the retrieved information contributes to the prediction. We omit residual connections, layer normalization, and the MLP block to isolate the attention and value dynamics. We consider $H = 2$ for two-head Transformers for the theoretical analysis in this paper.

Training to acquire/enhance ICL. To study how ICL capability can be acquired through training, we follow the theoretical framework introduced by Garg et al. (2022) and subsequently adopted in a number of works (Akyürek et al., 2023; Bai et al., 2023; von Oswald et al., 2023; Zhang et al., 2024; Huang et al., 2024; Li et al., 2024a). The model parameters Ψ are trained on prompts and their corresponding query outputs across a collection of tasks, so that the trained model learns to make in-context predictions on new prompts. Specifically, for each $n \in [N]$, a task τ_n is drawn from a training task set $\mathcal{T}_{\mathrm{tr}} \subset \mathcal{T}$, and a prompt $P^{(n)}$ is generated according to that task. The model is trained to predict the query output by minimizing the empirical squared prediction loss

$$\min_{\Psi} \quad \mathcal{L}(\Psi) = \frac{1}{N}\sum_{n=1}^{N} \ell\big(F(\Psi; P^{(n)}), y_q^{(n)}\big), \quad \text{with } P^{(n)} \sim \tau_n \in \mathcal{T}_{\mathrm{tr}} \tag{4}$$

where ℓ is the loss function. Although the supervised prompt-level objective in (4) is a simplified abstraction of practical model training for large language models, it provides a controlled setting

for analyzing how gradient-based training can lead to ICL capability, and the obtained theoretical insights are verified in practical experiments. We consider the squared prediction loss $\ell(a, b) = \|a - b\|_2^2/2$, which is commonly adopted in theoretical analyses of ICL for regression-style formulations (Li et al., 2025; Chen et al., 2024a; Huang et al., 2025; Kim & Suzuki, 2024; Li et al., 2024b; Mahankali et al., 2024).

ICL generalization. Let $\widehat{\Psi}$ denote the learned parameters by solving (4). For all query and example inputs drawn from $\mathcal{D}$ and all tasks $\tau \in \mathcal{T}$, the *in-domain generalization* error is

$$\mathbb{E}_{x_q \sim \mathcal{D},\, \tau \in \mathcal{T}} \quad \ell\big(F(\Psi; P), y_q\big) \tag{5}$$

Note that it also evaluates the unseen tasks in $\mathcal{T} \setminus \mathcal{T}_{tr}$ that do not appear in training. In addition, if all query and example inputs drawn from a distribution $\mathcal{D}' \neq \mathcal{D}$, and the tasks are drawn from a family $\mathcal{T}' \neq \mathcal{T}$, the *out-of-domain generalization* error is

$$\mathbb{E}_{x_q \sim \mathcal{D}',\, \tau \in \mathcal{T}'} \quad \ell\big(F(\Psi; P), y_q\big) \tag{6}$$

3 Theoretical Resutls

3.1 Main Results and Insights

We analyze ICL on compositional tasks, where each input consists of two latent components, each with M possible patterns, and a task specifies a separate input–output mapping for each component. We study a two-head Transformer in which both the attention and value parameters are trained jointly. This provides a minimal setting for understanding how multiple attention heads learn to specialize across different components of a compositional task. We summarize our main insights informally below, with formal results presented in subsequent sections. While our analysis focuses on two components and two heads for clarity, the underlying insights extend naturally to multi-component, multi-head Transformers.

(I1) Training convergence and in-domain ICL generalization on unseen tasks. We prove finite-step convergence of the joint training dynamics and show that the learned model generalizes to unseen tasks, even when their complete component mappings have not appeared during training. It suffices that every individual input–output correspondence occurring in the task family is observed in at least one training task. Consequently, for the full compositional family of up to $(M!)^2$ tasks, only M appropriately chosen training tasks are sufficient to satisfy this coverage condition.

(I2) Complementary head specialization. We prove that training amplifies weak differences in the initialization so that the two heads specialize to different components. Each head learns to match context examples according to one input component and retrieve its corresponding output component. Under uniform sampling, the required context length scales linearly with M, up to logarithmic factors. In contrast, jointly matching both components leads to an M^2 context-length scaling under the dependence established in prior single-head ICL analysis Li et al. (2024a).

(I3) Out-of-domain ICL generalization on unseen pattern families. We further show that the trained model can perform ICL on new compositional tasks whose input and output patterns were never observed during training. These tasks retain the same two-component structure but are defined over new pattern families obtained through constrained transformations of the training pattern families. Given context examples from such a new task, the model can infer its component-wise mappings and correctly predict the query output without further training.

3.2 Data Model and Parameterization

Data representation and Tasks. We consider two finite sets of latent input patterns $\mathcal{M} \subset \mathbb{R}^{d_\mu}$, $\mathcal{E} \subset \mathbb{R}^{d_e}$, with corresponding perturbation pattern sets $\mathcal{Z}_\mu \subset \mathbb{R}^{d_\mu}$, $\mathcal{Z}_e \subset \mathbb{R}^{d_e}$, and two finite sets of latent output patterns $\mathcal{N} \subset \mathbb{R}^{d_\nu}$, $\mathcal{F} \subset \mathbb{R}^{d_f}$. The patterns in $\mathcal{M} \cup \mathcal{Z}_\mu$, $\mathcal{E} \cup \mathcal{Z}_e$, $\mathcal{N}$, and $\mathcal{F}$ are pairwise orthogonal within their respective ambient spaces, and all patterns have norm γ. For notational simplicity, we assume $|\mathcal{M}| = |\mathcal{E}| = |\mathcal{N}| = |\mathcal{F}| = M$, and $|\mathcal{Z}_\mu| = |\mathcal{Z}_e| = M_z$ with $M_z \geq c^2 M$. For each pattern family, we use U_r ($r \in \{\mu, e, \nu, f, z_\mu, z_e\}$) to denote the matrix whose columns are the corresponding latent patterns. For example, $U_\mu = [\mu^{(1)}, \ldots, \mu^{(M)}]$, $U_{z_\mu} = [z_\mu^{(1)}, \ldots, z_\mu^{(M_z)}]$, and similarly for the other pattern families.

Each input x_l, including the query x_q, is generated by independently and uniformly sampling $\mu_l \sim \mathrm{Unif}(\mathcal{M})$, $e_l \sim \mathrm{Unif}(\mathcal{E})$, $z_l^\mu \sim \mathrm{Unif}(\mathcal{Z}_\mu)$, $z_l^e \sim \mathrm{Unif}(\mathcal{Z}_e)$, together with independent perturbation scales $\kappa_\mu, \kappa_e \sim \mathrm{Unif}[-c, c]$, $c > 0$, and takes the form

$$x_l = D_x \begin{bmatrix} \mu_l + \kappa_\mu z_l^\mu \\ e_l + \kappa_e z_l^e \end{bmatrix}, \quad l \in [L] \cup \{q\}. \tag{7}$$

A **task** $\tau = (g_\mu^\tau, g_e^\tau)$ in $\mathcal{T}$ specifies the input-output associations for the two pattern types through $g_\mu^\tau : \mathcal{M} \to \mathcal{N}$, $g_e^\tau : \mathcal{E} \to \mathcal{F}$. Both mappings are bijections. Given input x_l in (7), and task τ, the corresponding output is

$$y_l = \begin{bmatrix} g_\mu^\tau(\mu_l) \\ g_e^\tau(e_l) \end{bmatrix}, \quad l \in [L] \cup \{q\}. \tag{8}$$

Here, $D_x \in \mathbb{R}^{d_x \times (d_\mu + d_e)}$ has orthonormal columns. The two mappings can vary independently across tasks, so the task family $\mathcal{T}$ can contain up to $(M!)^2$ distinct tasks.

Remark 1. Our data model builds on structured pattern models commonly used in theoretical analyses of feature learning and in-context learning (Li et al., 2023a; Huang et al., 2024; Li et al., 2024a). Prior analyses primarily consider a single group of input–output pattern associations. For example, Li et al. (2024a) studies ICL with a single input–output pattern mapping. We extend this setting to compositional mappings, where a task consists of multiple component-wise mappings whose outputs jointly determine the overall input–output relation. Such compositional structure is prevalent in language and structured reasoning, where complex transformations are formed from reusable constituent components (Lake & Baroni, 2018; Liu & Neubig, 2022; An et al., 2023; Abedsoltan et al., 2025). We focus on two component mappings and two attention heads in (40) as the minimal setting for characterizing how multi-head attention learns and specializes to such compositional structure.

Condition 1 (Component-wise correspondence coverage)**.** *Every input–output correspondence that occurs in $\mathcal{T}$ is observed in at least one training task. Specifically, for every $(\mu, \nu) \in \mathcal{M} \times \mathcal{N}$ such that $g_\mu^\tau(\mu) = \nu$ for some $\tau \in \mathcal{T}$, there exists $\hat{\tau} \in \mathcal{T}_{\mathrm{tr}}$ such that $g_\mu^{\hat{\tau}}(\mu) = \nu$. Similarly, for every $(e, f) \in \mathcal{E} \times \mathcal{F}$ such that $g_e^\tau(e) = f$ for some $\tau \in \mathcal{T}$, there exists $\tilde{\tau} \in \mathcal{T}_{\mathrm{tr}}$ such that $g_e^{\tilde{\tau}}(e) = f$.*

Remark 2. Condition 1 requires coverage only of individual input–output correspondences, rather than complete component mappings or their combinations. Thus, the training task set $\mathcal{T}_{\mathrm{tr}}$ can be substantially smaller than the full task family $\mathcal{T}$. For example, $\mathcal{T}$ can contain all $(M!)^2$ combinations of two bijections, while M appropriately chosen training tasks are sufficient to satisfy Condition 1. To see this, index the patterns as $\{\mu_i, \nu_i, e_i, f_i\}_{i=1}^M$. For the k-th training task, $k = 0, \ldots, M-1$, choose the two component mappings as

$$g_\mu^{(k)}(\mu_i) = \nu_{1+((i-1+k) \bmod M)}, \qquad g_e^{(k)}(e_i) = f_{1+((i-1+k) \bmod M)}. \tag{9}$$

These M tasks cover all M^2 input–output correspondences for each component.

An **out-of-domain task** $\tau' = (g_\mu^{\tau'}, g_e^{\tau'})$ is defined analogously over new pattern sets $\mathcal{M}'$, $\mathcal{E}'$, $\mathcal{N}'$, $\mathcal{F}'$, $\mathcal{Z}_\mu'$, and $\mathcal{Z}_e'$, where $g_\mu^{\tau'} : \mathcal{M}' \to \mathcal{N}'$, $g_e^{\tau'} : \mathcal{E}' \to \mathcal{F}'$ are bijections. Inputs follow the same perturbation model.

Training setting. We train both the attention and value parameters and consider $H = 2$ attention heads. We restrict the effective parameterization in Section 2 to

$$W_{\mathrm{att}}^h = \begin{bmatrix} W_h & 0 \\ 0 & 0 \end{bmatrix}, \quad W_{\mathrm{OV}}^h = [0 \quad V_h], \quad V_h = \begin{bmatrix} \alpha_h I_{d_\nu} & 0 \\ 0 & \beta_h I_{d_f} \end{bmatrix} \in \mathbb{R}^{(d_\nu + d_f) \times (d_\nu + d_f)}, \tag{10}$$

where $W_h \in \mathbb{R}^{d_x \times d_x}$, and the two diagonal blocks of V_h correspond to the two output components. Consistent with structured parameterizations commonly used in theoretical analyses of ICL (Zhang et al., 2025; Li et al., 2024a; Kim & Suzuki, 2024; Sander et al., 2024; Chen et al., 2024a), the zero blocks separate input-based attention from output retrieval. Specifically, the attention weights are determined by the input features of the query and context examples, while the value mapping retrieves output information from the attended examples. The trainable coefficients α_h and β_h determine the contributions of the two output components. Let $\boldsymbol{\alpha} := (\alpha_1, \alpha_2)^\top \in \mathcal{V}$, and $\boldsymbol{\beta} := (\beta_1, \beta_2)^\top \in \mathcal{V}$. We constrain

$$\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{V} := \{w \in \mathbb{R}^2 : w_1 \geq 0,\ w_2 \geq 0,\ w_1 + w_2 \leq 1\}, \tag{11}$$

which provides a normalization of the value parameters and controls their scale during training.

Coupled training dynamics. The attention and value parameters, $\{W_1, W_2\}$ and $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$, are updated jointly using stochastic gradient descent with step size η_{att} and η_{v}, respectively. Note that the value parameters are fixed and only the attention parameters are learned (Li et al., 2024a; Huang et al., 2024; Shen et al., 2025) in existing ICL analysis, while our analysis will characterize their coupled evolution and show how the value parameters become specialized while the two attention heads increasingly concentrate on their respective input components.

Initialization. We use a structured symmetry-breaking initialization so that the attention heads have weak complementary preferences over the input components. Define the embedded pattern matrices

$$P_\mu = D_x \begin{bmatrix} U_\mu \\ \mathbf{0} \end{bmatrix}, \qquad P_e = D_x \begin{bmatrix} \mathbf{0} \\ U_e \end{bmatrix}, \qquad P_{z_\mu} = D_x \begin{bmatrix} U_{z_\mu} \\ \mathbf{0} \end{bmatrix}, \qquad P_{z_e} = D_x \begin{bmatrix} \mathbf{0} \\ U_{z_e} \end{bmatrix}, \tag{12}$$

where $P_\mu, P_e \in \mathbb{R}^{d_x \times M}$ and $P_{z_\mu}, P_{z_e} \in \mathbb{R}^{d_x \times M_z}$. For any $\theta > 0$, we initialize

$$W_1(0) = \frac{\theta}{\gamma^4}\left(P_\mu P_\mu^\top + P_{z_\mu} P_{z_\mu}^\top\right), \quad W_2(0) = \frac{\theta}{\gamma^4}\left(P_e P_e^\top + P_{z_e} P_{z_e}^\top\right), \quad \boldsymbol{\alpha}(0) = \boldsymbol{\beta}(0) = \mathbf{0}, \tag{13}$$

Here $\theta > 0$ controls the strength of the initial asymmetry. The initialization distinguishes only the two input-component subspaces and does not encode individual pattern correspondences or task-dependent mappings. Our analysis next will show that this weak initial preference is amplified through training into complementary head specialization. Exact structured initialization is not required, and the guarantees remain valid under sufficiently small Gaussian perturbations of $W_1(0)$ and $W_2(0)$, see Corollary 1 in Appendix.

3.3 Formal Theoretical Results

We now present our main theoretical results. We first establish finite-time convergence and compositional generalization to unseen combinations of component mappings (Theorem 1). We then characterize how the learned attention heads specialize to different input components (Proposition 1), and finally show the resulting model generalizes to out-of-domain pattern families (Theorem 2).

Theorem 1 (Convergence and compositional generalization)**.** *Suppose Condition 1 holds and the model is initialized according to (13). Let $M \geq 2$, $\theta > 0$, $c \in [0, 1/2]$, $0 < \epsilon \leq \gamma^2$, and $0 < \delta < 1$. Let $B \gtrsim \max\{M, \epsilon^{-2}\}\log\left(\frac{CM^2T}{\delta}\right)$, $\eta_{\text{att}} \lesssim \frac{\eta_{\text{v}}}{\gamma^4 \log(\frac{CM}{\eta_{\text{v}}\gamma^2})}$, $\eta_{\text{v}} \in (0, \frac{1}{2\gamma^2}]$, and*

$$L \gtrsim M\log\left(\frac{CM^2BT}{\delta}\right),\ T = O\left(\frac{M^2}{\eta_v\gamma^2}\log\frac{CM}{\eta_v\gamma^2} + \frac{M}{\eta_{\text{att}}\gamma^6}\left(\max\{(M-1)e^{-\frac{7\theta}{8}} - 1, 0\} + \frac{1}{\epsilon}\right)\right) \tag{14}$$

Then with probability at least $1 - \delta$ over the $N = BT$ training prompts, the following hold:

(i) ***Training convergence.*** *The training loss satisfies $\mathcal{L}(\widehat{\Psi}) \leq 2\gamma^2\epsilon$.*

(ii) ***Compositional generalization.*** *For every task $\tau \in \mathcal{T}$, including $\tau \notin \mathcal{T}_{\text{tr}}$, and an in-domain prompt P generated under task τ with query output y_q, the learned model satisfies*

$$\frac{1}{2}\|F(\widehat{\Psi}; P) - y_q\|_2^2 \leq 2\gamma^2\epsilon. \tag{15}$$

where $\widehat{\Psi} = \Psi(T)$ is the returned model after T iterations.

Remark 3. Theorem 1 establishes finite-time convergence to arbitrarily small error, with the number of training iterations scaling linearly with the target accuracy $1/\epsilon$. The bound on number iterations decreases with a higher bias level, θ, since attention is closer to the convergence. More importantly, the generalization guarantee does not require the complete test task $\tau = (g_\mu^\tau, g_e^\tau)$ to have appeared during training. Condition 1 requires only component-wise coverage: g_μ^τ and g_e^τ may appear in two different training tasks. The component-wise coverage can be achieved using only M appropriately chosen training tasks, compared with $|\mathcal{T}|$ possible tasks in the full family. We next show that this compositional generalization is accompanied by complementary attention head specialization.

Proposition 1 (Complementary head specialization)**.** *Under the assumptions of Theorem 1 and T satisfies (**??**), let $W_1(T)$, $W_2(T)$, $\boldsymbol{\alpha}(T)$, $\boldsymbol{\beta}(T)$ denote the learned weights. Then the following holds,*

(i) ***Attention specialization****. Head 1 (or 2) attends to context tokens that have the same μ (or e) pattern as the query, i.e.,*

$$\Big(1-\sum_{\ell\in\mathcal{N}_\mu^*}\mathrm{softmax}\left(x_\ell^\top W_1(T)x_q\right)\Big)^2<\epsilon,\quad \Big(1-\sum_{\ell\in\mathcal{N}_e^*}\mathrm{softmax}\left(x_\ell^\top W_2(T)x_q\right)\Big)^2<\epsilon \tag{16}$$

holds for any prompt P, where $\mathcal{N}_\mu^*:=\{\ell\in[L]:\mu_\ell=\mu_q\}$, $\mathcal{N}_e^*:=\{\ell\in[L]:e_\ell=e_q\}$.

(ii) ***Attention-matrix specialization****: let $\Phi_r^h(t)$ denote the* component-wise attention margin

$$\Phi_r^h(t):=\min_{i\in[M]}\min_{j\neq i}\left[\left(P_r^\top W_h(t)P_r\right)_{i,i}-\left(P_r^\top W_h(t)P_r\right)_{j,i}\right],\qquad h\in\{1,2\},r\in\{\mu,e\} \tag{17}$$

It holds that

$$\Phi_\mu^1(T_\epsilon)\geq\log\frac{M-1}{\sqrt{\epsilon}}-O(1),\qquad \Phi_e^2(T_\epsilon)\geq\log\frac{M-1}{\sqrt{\epsilon}}-O(1), \tag{18}$$

whereas the centered component-wise score magnitude satisfy

$$\left\|H_M P_e^\top W_1(T_\epsilon)P_e H_M\right\|_{\max}\leq O(1),\quad \left\|H_M P_\mu^\top W_2(T_\epsilon)P_\mu H_M\right\|_{\max}\leq O(1) \tag{19}$$

with $H_M:=I-\frac{1}{M}\mathbf{1}\mathbf{1}^\top$, and $\|U\|_{\max}=\max_{i,j}|U_{i,j}|$ for any matrix U.

(iii) ***Value specialization****. The value parameters specialize complementarily as*

$$\boldsymbol{\alpha}(T)=(1,0)^\intercal,\quad \boldsymbol{\beta}(T)=(0,1)^\intercal \tag{20}$$

Remark 4. Proposition 1 characterizes complementary head specialization. Head 1 attends to context examples that have the same μ pattern as the query, as shown in (16). Correspondingly, $P_\mu^\top W_1 P_\mu$ develops an increasingly large separation between matching and non-matching patterns as in (18) while the pattern-dependent variation of $P_e^\top W_1 P_e$ remains bounded as in (19). Head 2 exhibits the complementary behavior for the e-component. The value parameters further specialize as in (20), so that the two heads retrieve the corresponding output components.

Remark 5 (Implication for context length). The component-wise specialization helps explain why the context-length requirement in (**??**) scales linearly with M, up to logarithmic factors. a context example matches the μ (or e) component of the query with probability $1/M$. By comparison, Li et al. (2024a) establishes an inverse dependence on the fraction α' of examples that match the query pattern in a single-component setting. If the two components here were instead treated jointly, the corresponding probability of matching the joint (μ,e) pattern would be $1/M^2$, suggesting that the required context length in the order of M^2 by the scaling in Li et al. (2024a). While the assumptions of Li et al. (2024a) do not directly cover this joint-pattern construction, this comparison illustrates that the reduction of the context-length is enabled by component-wise head specialization.

Out-of-domain pattern shifts. We next study whether the learned model $\hat{\Psi}$ beyond the pattern families observed during training. We next consider the OOD pattern families $\mathcal{M}',\mathcal{E}',\mathcal{N}',\mathcal{F}',\mathcal{Z}'_\mu$ and $\mathcal{Z}'_e$ introduced in Section 2. Let $U'_\mu,U'_e,U'_\nu,U'_f,U'_{z_\mu},U'_{z_e}$ denote their corresponding pattern matrices, respectively. We consider structured OOD shifts in which each OOD pattern is a linear combination of patterns from the corresponding training family. Specifically, $U'_r=U_rC_r$, $r\in\{\mu,e,\nu,f,z_\mu,z_e\}$, where $C_r\in\mathbb{R}^{M\times M'}$ for $r\in\{\mu,e,\nu,f\}$, and $C_r\in\mathbb{R}^{M_z\times M'_z}$ for $r\in\{z_\mu,z_e\}$.

We impose mild boundedness and normalization conditions on these transformations. For the input patterns $r\in\{\mu,e\}$, each column c_i^r of C_r satisfies $\|c_i^r\|_2=1$, $\|c_i^r\|_1=O(1)$, $\mathbf{1}^\top c_i^r=\mathbf{1}^\top c_j^r$, for all i,j. For the output patterns $r\in\{\nu,f\}$, we only require $\|C_r\|_{\mathrm{op}}=O(1)$. For the perturbation patterns $r\in\{z_\mu,z_e\}$, $\|c_i^r\|_1=O(1)$, and $\|c_i^r\|_2=O(1)$.

Theorem 2 (Out-of-domain generalization)**.** *Suppose the assumptions of Theorem 1 hold, and let $\widehat{\Psi}=\Psi(T)$ be the learned parameters after T iterations. Consider an OOD task $\tau'\in\mathcal{T}'$ whose pattern families satisfy the conditions above. Then with high probability, for every OOD prompt P' generated under τ' with query output y'_q, the learned model satisfies*

$$\frac{1}{2}\left\|F(\widehat{\Psi};P')-y'_q\right\|_2^2\leq C_{\mathrm{ood}}\gamma^2(M-1)^{2\rho_{\mathrm{ood}}}\left(\frac{\epsilon}{(1-\sqrt{\epsilon})^2}\right)^{1-\rho_{\mathrm{ood}}}, \tag{21}$$

where $C_{\mathrm{ood}}>0$ is an constant and the OOD coherence $\rho_{\mathrm{ood}}:=\max_{r\in\{\mu,e\}}\max_{i\neq j}\left|(c_i^r)^\top c_j^r\right|$.

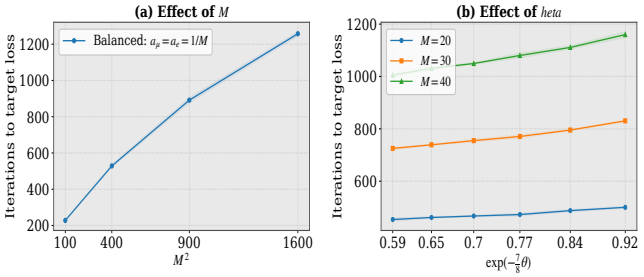

Figure 1: Number of iterations required against (a) M^2 and (b) $\exp(-7\theta/8)$.

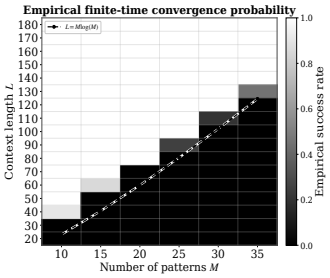

Figure 2: The required context length L versus M.

Theorem 2 shows that the specialization learned on in-domain pattern families extends to structured OOD pattern shifts, with the generalization error controlled by the OOD coherence ρ_{ood}. Smaller ρ_{ood} corresponds to better separation of the OOD input patterns and yields a stronger generalization guarantee. In particular, when the columns of C_μ and C_e are mutually orthogonal, $\rho_{\text{ood}} = 0$, and the OOD error remains of the same order as the in-domain error up to constant factors.

4 Numerical Experiments

4.1 Synthetic Experiments

We use the compositional data model in Section 3.2 and train the one-layer, two-head attention model as in (16). In-domain pattern families are constructed as mutually orthogonal coordinate directions. For the pattern families $\mathcal{M}, \mathcal{E}, \mathcal{N}, \mathcal{F}, \mathcal{Z}_\mu, \mathcal{Z}_e$, a separate orthonormal matrix is generated using QR decomposition of a random matrix and scaled by $\gamma = 2$. Then, the first $M = 20$ columns are chosen as the main pattern and the next $M_z = 30$ columns as perturbation patterns. The perturbation coefficients are sampled uniformly from $[-1/2, 1/2]$. Unless otherwise stated, we use context length $L = 200$, minibatch size $B = 128$, and initialization strength $\theta = 0.0$. The attention matrices are initialized according to (13), with additional Gaussian perturbations, and the attention and value parameters are trained jointly using SGD and projected SGD with step size $\eta_{\text{att}} = 0.2$ and $\eta_{\text{v}} = 0.4$, respectively. Training continues until the loss reaches 5×10^{-3} or the prescribed maximum number of iterations is reached.

Training iterations. Figure 3(a) reports the number of training iterations required to reach the target loss as M changes. The observed iteration complexity increases approximately linearly with M^2, verifying our theoretical bound in (**??**). Figure 3(b) shows the number of iterations required is indeed linear in $\exp(-7\theta/8)$, where θ is strength of the initial bias, as predicted by (**??**).

Context-length scaling. For each pair (M, L), we independently train the model for 30 trials, each for 5000 iterations. A trial is considered successful if the final training loss is below 0.01, and unsuccessful otherwise. Figure 2 shows the empirical success rate for different values of M and L. The transition boundary between successful and unsuccessful training scales approximately linearly with M, empirically supporting context-length requirement in (**??**) and Remark 5.

Head specialization. Figure 1(a) shows the total softmax attention mass assigned to context examples that share either the query's μ-pattern or its e-pattern. During training, Head 1 increasingly concentrates its attention on examples sharing the query's μ-pattern, whereas Head 2 concentrates on examples sharing its e-pattern. This verifies Proposition 1(i). Figure 1(b) shows that the specialized component-wise attention margins Φ_μ^1 and Φ_e^2 increase steadily during training, while the centered component-wise score magnitudes $\|H_M P_e^\top W_1 P_e H_M\|_{\max}$ and $\|H_M P_\mu^\top W_2 P_\mu H_M\|_{\max}$ remain comparatively small, consistent with (18) and (19) of Proposition 1(ii).

OOD ICL generalization. We control the OOD coherence ρ_{ood} by constructing $M' = M - 1$ OOD patterns as $\mu_j' = \sqrt{1-\rho_{\text{ood}}}\,\mu_j + \sqrt{\rho_{\text{ood}}}\,\mu_M$, $j \in [M-1]$, and construct the OOD e' patterns analogously. For the corresponding perturbation patterns, we set $z'_{\mu,j} = \frac{1}{\sqrt{2}}(z_{\mu,j} + z_{\mu,M})$,

$z'_{e,j} = \frac{1}{\sqrt{2}}(z_{e,j} + z_{e,M})$. Figure 4 shows that the logarithm of the OOD prediction error increases approximately linearly with ρ_{ood}, consistent with Theorem 2.

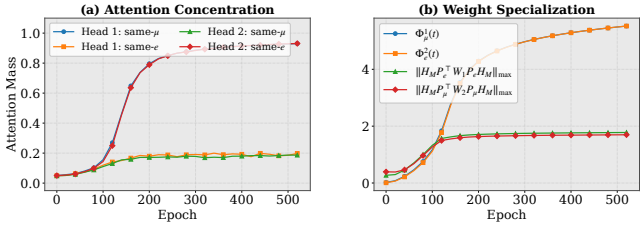

Figure 3: (a)Attention concentration emergence during training. (b) Weight specialization during training

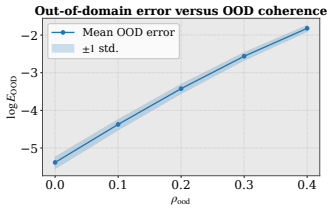

Figure 4: OOD prediction error versus OOD coherence ρ_{ood}.

4.2 Experiments on real-world image features

To evaluate head specialization in real-world data, we construct compositional ICL tasks using CIFAR-10 (Krizhevsky & Hinton, 2009) and DTD (Cimpoi et al., 2014). We select subsets from the two datasets and extract their features using a frozen pretrained ResNet-18 (He et al., 2015). Each input is formed by concatenating a CIFAR-10 embedding and a DTD embedding, corresponding to the μ and e components, respectively, and its output contains the labels associated with the two components. For each prompt, we randomly permute the class-to-label mappings, while applying each mapping within the prompt. Thus, examples from the same class share the same label within a prompt, but the label associated with a class can vary across prompts. This requires the model to infer the current mappings from the context examples.

We train a single-layer, two-head Transformer. Figure 5(a) shows the attention concentration of each head on context examples matching the query in the μ or e component. During training, Head 1 increasingly concentrates on the matching e-patterns, while Head 2 concentrates on the matching μ-patterns. In Figure 5(b), we average the training embeddings within each class and use the resulting class-wise mean embeddings as the empirical patterns, yielding 10 μ-patterns for CIFAR-10 and 10 e-patterns for DTD. Figure 5(b) shows the component-wise attention margins. The margins for the aligned head–component pairs, Φ_e^1 and Φ_μ^2, increase during training, whereas those for the misaligned pairs, Φ_μ^1 and Φ_e^2, decrease. These verify the complementary head specialization.

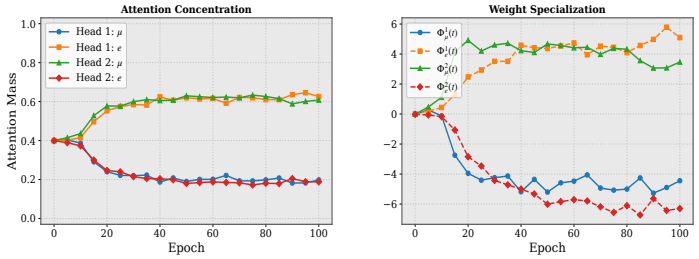

Figure 5: Head specialization on ICL tasks constructed from CIFAR-10 and DTD features. (a) Attention mass by each head to context examples matching the query's μ- or e-component class. (b) Component-wise attention margins for the aligned (Φ_e^1 and Φ_μ^2) and misaligned (Φ_μ^1 and Φ_e^2) pairs.

5 Conclusion

We studied how multi-head Transformers develop complementary specialization for in-context learning. Our theoretical analysis shows that joint training of attention and value parameters drives different heads to specialize in distinct input components, enabling generalization to unseen task combinations and structured distribution shifts. Experiments on synthetic and real-world data support the predicted training dynamics and specialization. Our results provide a theoretical explanation for how specialized attention heads emerge through training and contribute to ICL generalization.

AI USE STATEMENT

In this work, we used generative AI tools only for language polishing and improving the clarity of the writing. We did not use generative AI tools for research ideation, theoretical analysis, proofs, experiments, data analysis, or other aspects of the research. All AI-assisted edits were reviewed and verified by the authors. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical and does not involve human subjects or the collection of sensitive or private data. The empirical evaluations use publicly available benchmark datasets. We are not aware of any specific ethical concerns arising from the methods or experiments presented in this work.

REPRODUCIBILITY STATEMENT

We provide the complete assumptions and theoretical setup in the main text, with proofs of all theoretical results in the appendix. The experimental section and appendix describe the data generation procedures, model configurations, training settings, and real-data preprocessing used in our experiments.

REFERENCES

Amirhesam Abedsoltan, Huaqing Zhang, Kaiyue Wen, Hongzhou Lin, Jingzhao Zhang, and Mikhail Belkin. Task generalization with autoregressive compositional structure: Can learning from d tasks generalize to d^t tasks? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 154–173. PMLR, 2025.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8ed3d610ea4b68e7afb30ea7d01422c6-Abstract-Conference.html.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.15661.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11027–11052. Association for Computational Linguistics, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.04637.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0561738a239a995c8cd2ef0e50cfa4fd-Abstract-Conference.html.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality (extended abstract). In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, pp. 4573–4573, 2024a. URL https://proceedings.mlr.press/v247/siyu24a.html.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7aae9e3ec211249e05bd07271a6b1441-Abstract-Conference.html.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function lipschitzness. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a8633d27d782f66fe660c2fb4bae446e-Abstract-Conference.html.

Yingqian Cui, Jie Ren, Pengfei He, Hui Liu, Jiliang Tang, and Yue Xing. Superiority of multi-head attention: A theoretical study in shallow transformers in in-context linear regression. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 937–945. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/cui25a.html.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/75b0edb869e2cd509d64d0e8ff446bc1-Abstract-Conference.html.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2208.01066.

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 22686–22742, 2025. URL https://proceedings.mlr.press/v267/he25q.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

Jianhao Huang, Zixuan Wang, and Jason Lee. Transformers learn to implement multi-step gradient descent with chain of thought. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/da3d4d2e9b37f78ec3e7d0428c9b819a-Abstract-Conference.html.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 19660–19722, 2024. URL https://proceedings.mlr.press/v235/huang24d.html.

Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 37822–37836. Curran Associates, Inc., 2022. doi: 10.52202/068431-2741. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f69707de866eb0805683d3521756b73f-Paper-Conference.pdf.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 24527–24561, 2024. URL https://proceedings.mlr.press/v235/kim24af.html.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 2018.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=jClGv3Qjhb.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *Proceedings of the 41st International Conference on Machine Learning*, pp. 28734–28783, 2024a. URL https://proceedings.mlr.press/v235/li24bn.html.

Hongkang Li, Songtao Lu, Pin-Yu Chen, Xiaodong Cui, and Meng Wang. Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b295b3a940706f431076c86b78907757-Abstract-Conference.html.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. *arXiv preprint arXiv:2303.04245*, 2023b.

Zihao Li, Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason M. Klusowski, Jianqing Fan, and Mengdi Wang. One-layer transformer provably learns one-nearest neighbor in context. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/95827e011b9e899f189a01fe2f4ef316-Abstract-Conference.html.

Emmy Liu and Graham Neubig. Are representations built from the ground up? an empirical examination of local composition in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9053–9073. Association for Computational Linguistics, 2022.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8p3fu56lKc.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 38018–38070, 2024. URL https://proceedings.mlr.press/v235/nichani24a.html.

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, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html. Transformer Circuits technical report; arXiv:2209.11895.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. *arXiv preprint arXiv:2306.03435*, 2023.

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

Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6916–6932, 2024. doi: 10.18653/v1/2024.findings-acl.412. URL https://aclanthology.org/2024.findings-acl.412/.

Michael Eli Sander, Raja Giryes, Taiji Suzuki, Mathieu Blondel, and Gabriel Peyré. How do transformers perform in-context autoregressive learning ? 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. 43235–43254. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sander24a.html.

Wei Shen, Ruida Zhou, Jing Yang, and Cong Shen. On the training convergence of transformers for in-context classification of gaussian mixtures. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=iSyB2yYaMx.

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 *Proceedings of the 41st International Conference on Machine Learning*, pp. 45637–45662, 2024. URL https://proceedings.mlr.press/v235/singh24c.html.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *CoRR*, 2023.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arXiv:2305.16380*, 2023a.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Shaolei Du. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. In *Conference on Parsimony and Learning (Recent Spotlight Track)*, 2023b.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, Jo ao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 35151–35174, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9bfa0c155653e24120760a5ead819376-Abstract-Conference.html.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 72428–72461, 2025. URL https://proceedings.mlr.press/v267/yin25e.html.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL https://jmlr.org/papers/v25/23-1042.html.

Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew M. Saxe. Training dynamics of in-context learning in linear attention. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 76047–76087, 2025. URL https://proceedings.mlr.press/v267/zhang25br.html.

A Appendix

A.1 Noisy initialization corollary proof

Corollary 1 (Robustness to Gaussian attention initialization)**.** *Suppose the assumptions of Theorem 1 hold, except that*

$$\begin{aligned} W_1(0) &= \frac{\theta}{\gamma^4}\big(P_\mu P_\mu^\top + P_{z_\mu} P_{z_\mu}^\top\big) + \widetilde{W}_1, \\ W_2(0) &= \frac{\theta}{\gamma^4}\big(P_e P_e^\top + P_{z_e} P_{z_e}^\top\big) + \widetilde{W}_2, \end{aligned} \tag{22}$$

where the matrices $\widetilde{W}_1, \widetilde{W}_2$ are independent of the training data and have i.i.d. entries

$$(\widetilde{W}_h)_{ij} \sim \mathcal{N}\left(0, \frac{\sigma^2}{d_x}\right). \tag{23}$$

If

$$\sigma \le \frac{\sqrt{d_x}}{C\gamma^2(1+c)^2\sqrt{\log\big(C(M+M_z)^2/\delta\big)}} \min\left\{c_0\theta,\, c_{\mathrm{wrong}},\, \frac{c_{\mathrm{leak}}}{C_{\mathrm{prod}}},\, \frac{c_{\mathrm{sync}}}{C_{\mathrm{hand}}M}\right\}, \tag{24}$$

then, up to changes in numerical constants, the conclusions of Theorem 1 hold with probability at least $1-\delta$.

Proof. For any pair of dictionary patterns u, v,

$$u^\top \widetilde{W}_h v \sim \mathcal{N}\left(0, \frac{\sigma^2}{d_x}\|u\|_2^2\|v\|_2^2\right). \tag{25}$$

Hence, using the dictionary normalization and $|\kappa| \le c$, a Gaussian tail bound and a union bound over the $C(M+M_z)^2$ relevant probed coordinates give, with probability at least $1-\delta/2$,

$$\max_{\text{probes}} \left|x_\ell^\top \widetilde{W}_h x_q\right| \le C\frac{\sigma\gamma^2(1+c)^2}{\sqrt{d_x}}\sqrt{\log\frac{C(M+M_z)^2}{\delta}}. \tag{26}$$

Under equation 24, this perturbation fits simultaneously inside the initial score, wrong-class, leakage, and exchange-synchronization budgets used in the Phase-I handoff. Thus all initialization inequalities in the proof remain valid after changing only numerical constants, and the Phase-I and Phase-II arguments proceed unchanged. Applying the convergence result with failure probability $\delta/2$ for the fresh batches and taking a union bound with equation 26 proves the claim. □

A.2 Experiments Details

A.3 Real-data ICL experiment

Data and factor construction. We construct a real-data version of the ICL task in which each input is

$$x = [\mu, e].$$

Pattern μ is obtained from CIFAR-10 images and the factor e from DTD texture images. Both datasets are encoded with an ImageNet-pretrained ResNet-18 whose final classification layer is removed. The ResNet parameters are frozen throughout the experiment. For each channel separately, every feature coordinate is standardized using the mean and standard deviation computed from the corresponding training split, and the same transformation is then applied to the evaluation split. Thus,

$$\mu, e \in \mathbb{R}^{512}.$$

For each selected semantic class, we retain at most 300 training examples and 100 evaluation examples. A semantic class is represented by a collection of real image embeddings rather than by a single prototype. Hence, two occurrences of the same class generally correspond to different feature vectors. In this experiment, the repeated-pattern assumption of the synthetic model is therefore replaced by semantic class agreement.

Unless otherwise stated, we use $M_\mu = M_e = 10$ semantic classes. A query is formed by independently sampling a μ-class c_μ^q and an e-class c_e^q. For each prompt, two independent random permutations

$$g_\tau : [M_\mu] \to [M_\mu], \qquad h_\tau : [M_e] \to [M_e]$$

are sampled and define the task associated with that prompt. A context token with semantic classes $(c_{\mu,\ell}, c_{e,\ell})$ is

$$p_\ell = [\mu_\ell,\, e_\ell,\, \mathrm{onehot}(g_\tau(c_{\mu,\ell}))\,,\, \mathrm{onehot}(h_\tau(c_{e,\ell}))]\,,$$

whereas the query token contains no output label,

$$p_q = [\mu_q, e_q, 0, 0].$$

The prediction target is

$$\left(g_\tau(c_\mu^q),\, h_\tau(c_e^q)\right).$$

Because g_τ and h_τ are independently resampled for every prompt, the target cannot be inferred from the query representation alone.

The context contains three types of examples. The first shares the query's μ-class but has a different e-class, the second shares the query's e-class but has a different μ-class, and the remaining examples match neither factor. In the reported runs, approximately one quarter of the context is assigned to each of the first two groups and the remaining positions are used as distractors. The positions are randomly permuted within each prompt. With the default setting $L = 30$, this gives 8 μ-aligned examples, 8 e-aligned examples, and 14 distractors.

Model and training. The model contains one Transformer layer with two attention heads. We use $d_{\mathrm{model}} = 128$, so each head has dimension $d_h = 64$. Query and key projections receive only the real-valued factor features $[\mu, e]$. The mapped context labels are excluded from the query/key stream and enter only through the value/output stream. The input projection used by the query/key branch is bias-free, so the resulting attention score remains a bilinear function of the factor representations. Attention logits are scaled by $1/\sqrt{d_h}$, and the final query token is masked from attending to itself. The network uses neither LayerNorm nor residual connections. The attention output is passed through a two-layer MLP with expansion ratio 2, GELU activation, and zero dropout. Two output classifiers separately predict $g_\tau(c_\mu^q)$ and $h_\tau(c_e^q)$.

Training minimizes the sum of the two cross-entropy losses,

$$\mathcal{L} = \mathcal{L}_\mu + \mathcal{L}_e.$$

Prompts are generated online, so each optimization step uses a fresh batch of independently sampled tasks and contexts. We train with AdamW using learning rate 10^{-4}, weight decay 10^{-6}, batch size 128, and gradient-norm clipping at 1. The reported runs use 100 epochs with 50 optimization steps per epoch. Evaluation is performed every 5 epochs on a fixed bank of 512 evaluation prompts. This bank is generated once and held fixed during training so that changes in the diagnostics are attributable to changes in the model rather than to a changing evaluation sample.

Attention specialization. For each head h, we measure the total query-to-context attention mass assigned to context examples that agree with the query in the μ- or e-factor:

$$p_h(t) = \mathbb{E}\left[\sum_{\ell:\, c_{\mu,\ell} = c_\mu^q} a_\ell^{(h)}(t)\right], \qquad q_h(t) = \mathbb{E}\left[\sum_{\ell:\, c_{e,\ell} = c_e^q} a_\ell^{(h)}(t)\right].$$

The expectation is approximated using the fixed evaluation prompt bank. These quantities directly track whether a head increasingly selects context examples aligned with the query in one of the two factors.

To examine the same behavior at the query/key parameter level, the training representations are grouped according to their semantic labels. Let S_i^μ denote the set of μ-representations from class i. For head h, define

$$C_{\mu,h}(i,j) = \frac{1}{|S_i^\mu||S_j^\mu|} \sum_{\mu_r \in S_i^\mu} \sum_{\mu_s \in S_j^\mu} \mu_r^\top B_{\mu,h} \mu_s,$$

Table 2: Hyperparameters for the real-data factorized ICL experiment.

Setting	Value
μ dataset	CIFAR-10
e dataset	DTD
Feature extractor	Frozen ImageNet-pretrained ResNet-18
Feature dimension per factor	512
Feature preprocessing	Coordinatewise standardization
M_μ, M_e	$10, 10$
Context length L	100
Transformer layers	1
Attention heads	2
d_{model}	128
Head dimension d_h	64
MLP expansion ratio	2
LayerNorm	None
Residual connections	None
Dropout	0
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	10^{-6}
Batch size	128
Gradient clipping	1
Training epochs	100
Steps per epoch	50
Evaluation interval	5 epochs
Maximum training examples/class	300
Maximum evaluation examples/class	100

where

$$B_{\mu,h} = A_\mu^\top (W_K^{(h)})^\top W_Q^{(h)} A_\mu,$$

and A_μ denotes the μ-block of the learned input projection. The matrix $C_{e,h}$ is defined analogously using

$$B_{e,h} = A_e^\top (W_K^{(h)})^\top W_Q^{(h)} A_e.$$

Thus, $C_{\mu,h}$ and $C_{e,h}$ are empirical class-level analogues of $P_\mu^\top W_h P_\mu$ and $P_e^\top W_h P_e$, respectively. The trajectory calculations use fixed class-grouped training pools at every checkpoint, so their support does not change during training. For the initialization-versus-final heatmaps, the same set of training patterns is used at both times.

We additionally remove class-independent row and column components with

$$H = I - \frac{1}{M}\mathbf{1}\mathbf{1}^\top.$$

For each head, we track

$$\|HC_{\mu,h}(t)H\|_{\max}, \qquad \|HC_{e,h}(t)H\|_{\max},$$

and

$$\|HC_{\mu,h}(t)H\|_{\min}, \qquad \|HC_{e,h}(t)H\|_{\min},$$

where

$$\|A\|_{\max} = \max_{i,j} |A_{ij}|, \qquad \|A\|_{\min} = \min_{i,j} |A_{ij}|.$$

The centered quantities isolate the class-dependent part of the learned bilinear attention operator and are plotted together with the aligned-token attention masses.

A.4 Proof summary

Proof architecture. The proof has two phases. Phase I shows that the value parameters reach the specialized vertices while the attention matrices remain close enough to initialization to preserve

a positive KKT margin. At the handoff time T_1, we obtain five properties: value specialization, preferred-class maximality, a wrong-class score bound, leakage control, and cross-query synchronization. Phase II propagates these properties by a joint stopped induction. Under this induction, the clean preferred logits have positive drift, leakage remains bounded, and query classes remain synchronized. These facts reduce the dynamics to a scalar recurrence for the average clean log-odds. Solving this recurrence yields attention concentration, which is then converted to prediction error. The proof is written for head 1, the head-2 statements follow after exchanging the μ- and e-channels.

A.5 Setup and notation

We consider the two-head Transformer, data model, and simplex-constrained value parameters defined in the main paper. The proof is written for head 1; the corresponding statements for head 2 follow by exchanging the μ- and e-channels. The main and perturbation indices are mutually independent and uniform on $[M]$ and $[M_z]$, respectively.

Value constraint and sampling. The value coefficients belong to

$$\mathcal{V} = \{w \in \mathbb{R}^2 : w_1, w_2 \geq 0,\ w_1 + w_2 \leq 1\}, \tag{27}$$

and are updated by projected stochastic gradient descent on $\mathcal{V}$.

We consider a balanced-batch case. Every batch contains B/M queries of each μ-class and B/M queries of each e-class. For a fixed μ-class, the associated e-indices are uniform across the permutation blocks, and the symmetric statement holds after exchanging the two channels. The concentration result for this sampling scheme is given in Appendix A.20.

Initialization and clean attention coordinates. For a fixed $\theta > 0$, initialize

$$W_1(0) = \frac{\theta}{\gamma^4}\left(P_\mu P_\mu^\top + P_{z_\mu} P_{z_\mu}^\top\right), \qquad W_2(0) = \frac{\theta}{\gamma^4}\left(P_e P_e^\top + P_{z_e} P_{z_e}^\top\right), \qquad \alpha(0) = \beta(0) = 0. \tag{28}$$

By the dictionary normalization,

$$P_\mu^\top W_1(0) P_\mu = \theta I_M, \qquad P_e^\top W_2(0) P_e = \theta I_M, \tag{29}$$

while all other probed cross blocks vanish at initialization.

For head 1, define the clean main-pattern score matrix

$$A(t) = P_\mu^\top W_1(t) P_\mu. \tag{30}$$

For query class q, let

$$p_q(t) = \frac{\exp(A_{q,q}(t))}{\sum_{s=1}^M \exp(A_{s,q}(t))}, \qquad a_q(t) = \log \frac{p_q(t)}{1 - p_q(t)}. \tag{31}$$

At initialization these quantities are independent of q, and

$$p(0) = \frac{e^\theta}{e^\theta + M - 1}, \qquad \frac{1 - p(0)}{M - 1} = \frac{1}{e^\theta + M - 1}. \tag{32}$$

The corresponding head-2 quantities are defined by exchanging μ and e.

Proof radii and constant convention. The initialization bias $\theta > 0$, feature scale $\gamma > 0$, and perturbation bound $c > 0$ are fixed model parameters. Throughout the proof, $C > 0$ denotes a numerical constant that may change from line to line and depends only on the fixed model parameters. We choose fixed radii

$$0 < c_{\mathrm{leak}} \ll c_{\mathrm{wrong}} \ll c_{\mathrm{sync}} \ll \theta, \tag{33}$$

where c_{leak} controls centered non-signal scores, c_{wrong} controls the spread among clean wrong-class scores, and c_{sync} controls differences between query classes. More specific constants are introduced locally when they are needed.

The fixed perturbation component in $W_h(0)$ is treated separately from learned leakage: c_{leak} controls the centered contribution of $W_h(t) - W_h(0)$, whereas the fixed initialization contribution is controlled through class-wise concentration.

Role of the initialization bias. The parameter θ selects the complementary specialization branch at initialization. We take $\theta > 0$ fixed independently of M. As $\theta \downarrow 0$, both the initial specialization margin and the conditioning used in Phase I degenerate, and the Phase-I convergence time increases.

A.6 Exact gradients and value dynamics

For prompt n and head $h \in \{1, 2\}$, define the model output

$$\widehat{y}_q^{(n)} = \sum_{h=1}^{2} V_h \sum_{\ell=1}^{L} \mathrm{softmax}((x_\ell^{(n)})^\top W_h x_q^{(n)}) y_\ell^{(n)}. \tag{34}$$

Let

$$z_{\ell,h}^{(n)} = (x_\ell^{(n)})^\top W_h x_q^{(n)}, \qquad \mathrm{softmax}(z_{\ell,h}^{(n)}) = \sigma_{\ell,h}^{(n)}(z_{\ell,h}^{(n)}) = \frac{\exp(z_{\ell,h}^{(n)})}{\sum_{r=1}^{L} \exp(z_{r,h}^{(n)})}. \tag{35}$$

And

$$u_h^{(n)} = \sum_{\ell=1}^{L} \sigma_{\ell,h}^{(n)} \nu_{a_\ell^{(n)}}, \qquad v_h^{(n)} = \sum_{\ell=1}^{L} \sigma_{\ell,h}^{(n)} f_{i_\ell^{(n)}}. \tag{36}$$

Equivalently, if

$$\pi_a^{(n,h)} = \sum_{\ell : a_\ell^{(n)} = a} \sigma_{\ell,h}^{(n)}, \qquad \rho_i^{(n,h)} = \sum_{\ell : i_\ell^{(n)} = i} \sigma_{\ell,h}^{(n)}, \tag{37}$$

then

$$u_h^{(n)} = \sum_{a=1}^{M} \pi_a^{(n,h)} \nu_a, \qquad v_h^{(n)} = \sum_{i=1}^{M} \rho_i^{(n,h)} f_i. \tag{38}$$

Write

$$\alpha = (\alpha_1, \alpha_2)^\top, \qquad \beta = (\beta_1, \beta_2)^\top, \tag{39}$$

and, in the orthogonal output dictionary coordinates, the model output is the sum of the two head outputs,

$$\widehat{y}_q^{(n)} = \begin{bmatrix} \alpha_1 u_1^{(n)} + \alpha_2 u_2^{(n)} \\ \beta_1 v_1^{(n)} + \beta_2 v_2^{(n)} \end{bmatrix}. \tag{40}$$

Define

$$r_\nu^{(n)} = \nu_{j_n} - \sum_{h=1}^{2} \alpha_h u_h^{(n)}, \qquad r_f^{(n)} = f_{i_n} - \sum_{h=1}^{2} \beta_h v_h^{(n)}. \tag{41}$$

Then, the single sample loss is

$$y_q^{(n)} - \widehat{y}_q^{(n)} = \begin{bmatrix} r_\nu^{(n)} \\ r_f^{(n)} \end{bmatrix}. \tag{42}$$

Thus the batch loss is

$$\mathcal{L}_t = \frac{1}{2B} \sum_{n \in \mathcal{B}_t} \left(\|r_\nu^{(n)}\|_2^2 + \|r_f^{(n)}\|_2^2 \right). \tag{43}$$

Lemma 1 (Exact gradients of the attention and value parameters)**.** *For $h \in \{1, 2\}$, define*

$$g_{a,h}^{(n)} = \gamma^{-2} \left\langle r_\nu^{(n)}, \nu_a - u_h^{(n)} \right\rangle, \qquad \widetilde{g}_{i,h}^{(n)} = \gamma^{-2} \left\langle r_f^{(n)}, f_i - v_h^{(n)} \right\rangle, \tag{44}$$

and

$$\Gamma_{\ell,h}^{(n)} = \alpha_h g_{a_\ell^{(n)},h}^{(n)} + \beta_h \widetilde{g}_{i_\ell^{(n)},h}^{(n)}. \tag{45}$$

Then the exact attention gradient is

$$\nabla_{W_h} \mathcal{L}_t = -\frac{\gamma^2}{B} \sum_{n \in \mathcal{B}_t} \sum_{\ell=1}^{L} \sigma_{\ell,h}^{(n)} \Gamma_{\ell,h}^{(n)} x_\ell^{(n)} (x_q^{(n)})^\top. \tag{46}$$

For the value coefficients,

$$\frac{\partial \mathcal{L}_t}{\partial \alpha_h} = -\frac{1}{B}\sum_{n\in\mathcal{B}_t}\left\langle r_\nu^{(n)}, u_h^{(n)}\right\rangle, \qquad \frac{\partial \mathcal{L}_t}{\partial \beta_h} = -\frac{1}{B}\sum_{n\in\mathcal{B}_t}\left\langle r_f^{(n)}, v_h^{(n)}\right\rangle. \tag{47}$$

Equivalently,

$$\nabla_\alpha \mathcal{L}_t = \gamma^2\left(G_\alpha(t)\alpha(t) - c_\alpha(t)\right), \qquad \nabla_\beta \mathcal{L}_t = \gamma^2\left(G_\beta(t)\beta(t) - c_\beta(t)\right), \tag{48}$$

where

$$G_\alpha(t) = \frac{1}{B}\sum_{n\in\mathcal{B}_t}\begin{bmatrix} \|u_1^{(n)}\|_2^2/\gamma^2 & \langle u_1^{(n)}, u_2^{(n)}\rangle/\gamma^2 \\ \langle u_1^{(n)}, u_2^{(n)}\rangle/\gamma^2 & \|u_2^{(n)}\|_2^2/\gamma^2 \end{bmatrix}, \tag{49}$$

$$c_\alpha(t) = \frac{1}{B}\sum_{n\in\mathcal{B}_t}\begin{bmatrix} \pi_{j_n}^{(n,1)} \\ \pi_{j_n}^{(n,2)} \end{bmatrix}, \tag{50}$$

and

$$G_\beta(t) = \frac{1}{B}\sum_{n\in\mathcal{B}_t}\begin{bmatrix} \|v_1^{(n)}\|_2^2/\gamma^2 & \langle v_1^{(n)}, v_2^{(n)}\rangle/\gamma^2 \\ \langle v_1^{(n)}, v_2^{(n)}\rangle/\gamma^2 & \|v_2^{(n)}\|_2^2/\gamma^2 \end{bmatrix}, \tag{51}$$

$$c_\beta(t) = \frac{1}{B}\sum_{n\in\mathcal{B}_t}\begin{bmatrix} \rho_{i_n}^{(n,1)} \\ \rho_{i_n}^{(n,2)} \end{bmatrix}. \tag{52}$$

Consequently, the projected value updates are

$$\alpha(t+1) = \Pi_{\mathcal{V}}\left(\alpha(t) - \eta_v\gamma^2\left[G_\alpha(t)\alpha(t) - c_\alpha(t)\right]\right), \tag{53}$$

$$\beta(t+1) = \Pi_{\mathcal{V}}\left(\beta(t) - \eta_v\gamma^2\left[G_\beta(t)\beta(t) - c_\beta(t)\right]\right). \tag{54}$$

Proof. Because the output dictionary is orthonormal, the squared prediction error can be evaluated in the two output blocks before applying the dictionary. This gives equation 43.

We first differentiate with respect to the value coefficients. The retrievals $u_h^{(n)}$ and $v_h^{(n)}$ do not depend on α or β. Therefore

$$\frac{\partial r_\nu^{(n)}}{\partial \alpha_h} = -u_h^{(n)}, \qquad \frac{\partial r_f^{(n)}}{\partial \beta_h} = -v_h^{(n)}. \tag{55}$$

Applying the chain rule to equation 43 yields

$$\frac{\partial \mathcal{L}_t}{\partial \alpha_h} = \frac{1}{B}\sum_{n\in\mathcal{B}_t}\left\langle r_\nu^{(n)}, \frac{\partial r_\nu^{(n)}}{\partial \alpha_h}\right\rangle = -\frac{1}{B}\sum_{n\in\mathcal{B}_t}\left\langle r_\nu^{(n)}, u_h^{(n)}\right\rangle, \tag{56}$$

and the same calculation in the f-block gives the β_h formula in equation 47.

Expanding the first expression in equation 47 gives

$$\frac{\partial \mathcal{L}_t}{\partial \alpha_h} = \frac{1}{B}\sum_{n\in\mathcal{B}_t}\left[\sum_{r=1}^{2}\alpha_r\left\langle u_r^{(n)}, u_h^{(n)}\right\rangle - \left\langle \nu_{j_n}, u_h^{(n)}\right\rangle\right]. \tag{57}$$

The output patterns are orthogonal and have squared norm γ^2. Using equation 38,

$$\left\langle \nu_{j_n}, u_h^{(n)}\right\rangle = \gamma^2\pi_{j_n}^{(n,h)}. \tag{58}$$

Dividing the retrieval inner products by γ^2 therefore gives

$$\nabla_\alpha \mathcal{L}_t = \gamma^2\left(G_\alpha(t)\alpha(t) - c_\alpha(t)\right). \tag{59}$$

The f-channel calculation is identical and gives the second identity in equation 48.

It remains to differentiate the attention parameters. For fixed prompt n and head h, the softmax derivative is

$$\frac{\partial \sigma_{r,h}^{(n)}}{\partial z_{\ell,h}^{(n)}} = \sigma_{r,h}^{(n)} \left(\mathbf{1}\{r = \ell\} - \sigma_{\ell,h}^{(n)} \right). \tag{60}$$

Hence

$$\frac{\partial u_h^{(n)}}{\partial z_{\ell,h}^{(n)}} = \sigma_{\ell,h}^{(n)} \left(\nu_{a_\ell^{(n)}} - u_h^{(n)} \right), \tag{61}$$

and similarly

$$\frac{\partial v_h^{(n)}}{\partial z_{\ell,h}^{(n)}} = \sigma_{\ell,h}^{(n)} \left(f_{i_\ell^{(n)}} - v_h^{(n)} \right). \tag{62}$$

Only head h's retrievals depend on $z_{\ell,h}^{(n)}$, so

$$\frac{\partial r_\nu^{(n)}}{\partial z_{\ell,h}^{(n)}} = -\alpha_h \sigma_{\ell,h}^{(n)} \left(\nu_{a_\ell^{(n)}} - u_h^{(n)} \right), \tag{63}$$

$$\frac{\partial r_f^{(n)}}{\partial z_{\ell,h}^{(n)}} = -\beta_h \sigma_{\ell,h}^{(n)} \left(f_{i_\ell^{(n)}} - v_h^{(n)} \right). \tag{64}$$

Applying the chain rule to the prompt loss gives

$$\frac{\partial}{\partial z_{\ell,h}^{(n)}} \frac{1}{2} \left(\|r_\nu^{(n)}\|_2^2 + \|r_f^{(n)}\|_2^2 \right) = -\gamma^2 \sigma_{\ell,h}^{(n)} \Gamma_{\ell,h}^{(n)}, \tag{65}$$

where $\Gamma_{\ell,h}^{(n)}$ is exactly equation 45.

Finally,

$$\frac{\partial z_{\ell,h}^{(n)}}{\partial W_h} = x_\ell^{(n)} (x_q^{(n)})^\top. \tag{66}$$

Combining equation 65 and equation 66, summing over the L context positions and the B prompts, and dividing by B proves equation 46. The projected updates equation 53–equation 54 then follow directly from projected gradient descent on the simplex $\mathcal{V}$. □

The matrices G_α, G_β and vectors c_α, c_β govern the projected value dynamics. They will be used in Phase I to establish contraction toward the specialized simplex vertices and to verify the corresponding KKT conditions.

A.7 Value convergence and handoff

Phase I drives the value coefficients to complementary simplex vertices while the attention parameters remain close enough to their initialization to preserve the value-gradient geometry. We first quantify the KKT margin and Gram conditioning induced by the complementary initialization. We then show that these properties persist until the value coefficients reach the specialized vertices. The resulting handoff conditions initialize the Phase-II bootstrap.

A.8 Initialization margin and Phase-I horizon

At the symmetric initialization, define

$$m_{\mathrm{KKT}}(0) := (1 - p(0)) \left(p(0) - \frac{1 - p(0)}{M - 1} \right). \tag{67}$$

Let $G_{\alpha,0}$ and $G_{\beta,0}$ denote the corresponding population Gram matrices at initialization, and define

$$\lambda_M := \lambda_{\min}(G_{\alpha,0}) = \lambda_{\min}(G_{\beta,0}). \tag{68}$$

The quantity $m_{\mathrm{KKT}}(0)$ measures the initial margin toward the specialized simplex vertices, while λ_M determines the contraction rate of the projected value dynamics.
Choose $c_0 > 0$, depending only on θ, so that

$$c_0\theta < \frac{1}{4}. \tag{69}$$

$$\max\left\{\left|e^{-2c_0\theta}\frac{1-c_0\theta}{1+c_0\theta}-1\right|, \left|e^{2c_0\theta}\frac{1+c_0\theta}{1-c_0\theta}-1\right|\right\} \le \frac{1-e^{-\theta}}{16(1+e^{\theta})}. \tag{70}$$

$$4c_0\theta + 2\log\frac{1+c_0\theta}{1-c_0\theta} \le \frac{\theta}{2}. \tag{71}$$

Such a choice exists because the left sides tend to zero with c_0.
Let

$$T_{\mathrm{pre}} := \min\left\{t \ge 0 : \left(1-\frac{\eta_v\gamma^2\lambda_M}{C_{\mathrm{Gram}}}\right)^t \le \frac{m_{\mathrm{KKT}}(0)\eta_v\gamma^2}{32}\right\}, \qquad T_1 := T_{\mathrm{pre}} + 1, \tag{72}$$

where C_{Gram} is the fixed conditioning constant.
The time T_{pre} brings the value coefficients into a small neighborhood of their target vertices; the additional update at T_1 projects them exactly onto those vertices.
Let $c_{\mathrm{wrong}}, c_{\mathrm{sync}}, c_{\mathrm{leak}} > 0$ be fixed.

Proposition 2 (Phase I handoff)**.** *Assume the balanced query design, $M \ge 2$, $M_z \ge Cc^2M$, and $0 < \eta_v\gamma^2 \le 1/2$. Different batches are independent and each batch is used once.*

Let

$$L, B \ge CM\log\left(\frac{C(M+M_z)^2B(T_1+1)}{\delta}\right). \tag{73}$$

Choose the numerical constant in equation 73 large enough that the relative class-weight concentration error is at most both $c_0\theta$ and a number whose contribution to a difference of two class log-normalizers is at most $c_{\mathrm{leak}}/(8C_{\mathrm{prod}})$.

Assume also that η_{att} is small enough that

$$C\eta_{\mathrm{att}}\gamma^6T_1\left[\frac{1}{M}+c^2\left(\frac{1}{M_z}+\frac{1}{M}\right)\right]^2 \le \min\left\{c_0\theta, \frac{c_{\mathrm{wrong}}}{8}, \frac{c_{\mathrm{leak}}}{8C_{\mathrm{prod}}}, \frac{\theta}{8}\right\}, \tag{74}$$

and

$$\begin{aligned} C\exp\left(\frac{C\eta_{\mathrm{att}}\gamma^6T_1}{M^2}\right)\Bigg[&\sqrt{\frac{\eta_{\mathrm{att}}^2\gamma^{12}T_1}{BM^3}\log\frac{C(M+M_z)^2M^2(T_1+1)}{\delta}} \\ &+\frac{\eta_{\mathrm{att}}\gamma^6}{BM}\log\frac{C(M+M_z)^2M^2(T_1+1)}{\delta}\Bigg] \\ \le\ &\frac{c_{\mathrm{sync}}}{C_{\mathrm{hand}}M}. \end{aligned} \tag{75}$$

Finally choose the fixed radii so that

$$2c_0\theta + \frac{c_{\mathrm{leak}}}{2C_{\mathrm{prod}}} \le \frac{\theta}{2}. \tag{76}$$

Then, with probability at least $1-\delta$, all of the following hold simultaneously.

1. Value specialization:
$$\alpha(T_1) = (1,0)^\top, \qquad \beta(T_1) = (0,1)^\top. \tag{77}$$
2. KKT and conditioning at the handoff: *the two α-KKT margins at $(1,0)^\top$ and the two β-KKT margins at $(0,1)^\top$ are at least $m_{\mathrm{KKT}}(0)/2$, and*
$$\frac{\lambda_M}{C_{\mathrm{Gram}}}I_2 \preceq G_\alpha(T_1), G_\beta(T_1) \preceq 2I_2. \tag{78}$$
Thus the specialized vertices are fixed points of the next projected value update as long as these margins persist.

3. Score tube and wrong-class spread: *for every sampled token and query used through T_1,*

$$\left|(D_x x_\ell)^\top (W_h(t) - W_h(0)) D_x x_q\right| \le c_0\theta, \qquad 0 \le t \le T_1,\ h = 1, 2. \tag{79}$$

Moreover, for $s, r \neq q$,

$$|A_{s,q}(T_1) - A_{r,q}(T_1)| \le \frac{c_{\mathrm{wrong}}}{4}, \qquad A_{q,q}(T_1) - A_{s,q}(T_1) \ge \theta - 2c_0\theta. \tag{80}$$

The head-2 analogue holds after exchanging μ and e.

4. Preferred-class maximality: *in every handoff prompt with μ-query class q, the full head-1 attention mass of class q is larger than that of every class $s \neq q$; in fact the log-ratio is at least $\theta/2$. The symmetric assertion holds for head 2 and the e-classes.*

5. Cross-query synchronization:

$$\max_{q,q'} \left| \frac{1 - p_q(T_1)}{p_q(T_1)} - \frac{1 - p_{q'}(T_1)}{p_{q'}(T_1)} \right| \le \frac{c_{\mathrm{sync}}}{16}, \tag{81}$$

and likewise for head 2 after exchanging the channels.

6. Leakage base case: *the contribution to every centered non-signal score from $W_h(T_1) - W_h(0)$ is at most $c_{\mathrm{leak}}/(8C_{\mathrm{prod}})$. The difference between the fixed-initialization nuisance corrections to any two main-class log-normalizers is at most $c_{\mathrm{leak}}/(8C_{\mathrm{prod}})$. Hence the total class-dependent non-signal correction to a preferred-versus-wrong logit difference is at most $c_{\mathrm{leak}}/(2C_{\mathrm{prod}})$.*

For fixed $\theta, c, c_{\mathrm{wrong}}, c_{\mathrm{sync}}, c_{\mathrm{leak}}$ the familiar simpler scaling

$$T_1 = O_\theta\left(\frac{M^2}{\eta_v \gamma^2} \log \frac{CM}{\eta_v \gamma^2}\right), \qquad \eta_{\mathrm{att}} \le \frac{\eta_v}{C\gamma^4 \log(CM/(\eta_v\gamma^2))} \left(1 + c^2 + \frac{c^2 M}{M_z}\right)^{-2} \tag{82}$$

with the fixed denominator constant chosen sufficiently large, together with equation 73 with a sufficiently large numerical constant, implies equation 74–equation 75.

A.9 Proof of the Phase-I handoff

A.9.1 One probability event

Pre-sample the fresh batches through the batch used at the handoff. For a fixed prompt and a fixed main class, the initialization class weight is the sum of independent bounded nonnegative variables. Conditional on the query, its mean is at least a fixed constant times L/M. Multiplicative Chernoff therefore gives, for every prescribed $0 < \rho < 1$,

$$\mathbb{P}\left(\left|\frac{\text{class weight}}{\mathbb{E}[\text{class weight} \mid x_q]} - 1\right| > \rho \,\middle|\, x_q\right) \le 2\exp\left(-\frac{\rho^2 L}{CM}\right). \tag{83}$$

Use equation 83 first with $\rho = c_0\theta$ and then with a fixed ρ small enough that

$$2\log\frac{1+\rho}{1-\rho} \le \frac{c_{\mathrm{leak}}}{8C_{\mathrm{prod}}}. \tag{84}$$

There are at most $CMB(T_1 + 1)$ relevant class weights. Thus equation 73 and a union bound make both estimates simultaneous.

The same argument, now for counts, gives simultaneously

$$\begin{aligned}
\frac{\#\{\text{tokens with a fixed main index}\}}{L} &\le \frac{C}{M}, \\
\frac{\#\{\text{queries with a fixed main index}\}}{B} &\le \frac{C}{M}, \\
\frac{\#\{\text{tokens with a fixed perturbation index}\}}{L} &\le C\left(\frac{1}{M_z} + \frac{1}{M}\right), \\
\frac{\#\{\text{queries with a fixed perturbation index}\}}{B} &\le C\left(\frac{1}{M_z} + \frac{1}{M}\right).
\end{aligned} \tag{85}$$

The union bound also includes the synchronization estimate proved in Step 5. We work on this single event for the rest of the proof. Its probability is at least $1 - \delta$ after splitting δ among the finite collections above.

A.9.2 KKT MARGINS AND GRAM CONDITIONING INSIDE THE SCORE TUBE

Lemma 2 (Phase-I KKT margin and Gram conditioning)**.** *Assume the Phase-I concentration event holds and, at some $t \le T_1$,*

$$\max_{n,\ell,h} \left| (x_\ell^{(n)})^\top \big(W_h(t) - W_h(0)\big) x_q^{(n)} \right| \le c_0\theta. \tag{86}$$

Then the two KKT margins for $\alpha^\star = (1,0)^\top$ and the two KKT margins for $\beta^\star = (0,1)^\top$ are at least

$$\frac{m_{\mathrm{KKT}}(0)}{2}.$$

Moreover,

$$\frac{\lambda_M}{C_{\mathrm{Gram}}} I_2 \preceq G_\alpha(t), G_\beta(t) \preceq 2I_2. \tag{87}$$

Proof. Assume temporarily that

$$\left| (D_x x_\ell)^\top (W_h(t) - W_h(0)) D_x x_q \right| \le c_0\theta. \tag{88}$$

Define

$$\mathcal{V} = \{w \in \mathbb{R}^2 : w_1, w_2 \ge 0,\ w_1 + w_2 \le 1\}. \tag{89}$$

For an arbitrary fixed $\theta > 0$, initialize

$$W_1(0) = \frac{\theta}{\gamma^4}\left(P_\mu P_\mu^\top + P_{z_\mu} P_{z_\mu}^\top\right), \qquad W_2(0) = \frac{\theta}{\gamma^4}\left(P_e P_e^\top + P_{z_e} P_{z_e}^\top\right), \qquad \alpha(0) = \beta(0) = 0, \tag{90}$$

The value coefficients are updated by projected stochastic gradient descent on $\mathcal{V}$. The two KKT margins at $(1,0)^\top$ are

$$(c_\alpha)_1 - (G_\alpha)_{11}, \qquad (c_\alpha)_1 - (c_\alpha)_2 - (G_\alpha)_{11} + (G_\alpha)_{12}. \tag{91}$$

Similarly, the two KKT margins at $(0,1)^\top$ are

$$(c_\beta)_2 - (G_\beta)_{22}, \qquad (c_\beta)_2 - (c_\beta)_1 - (G_\beta)_{22} + (G_\beta)_{12}. \tag{92}$$

Positivity of these quantities is equivalent to the strict KKT conditions at the corresponding vertex. Let p be a head-1 μ-class profile in a prompt and let q be the head-2 μ-class profile in that prompt. At balanced population initialization their profiles are respectively $(p(0), \frac{1-p(0)}{M-1}, \ldots, \frac{1-p(0)}{M-1})$, with the preferred coordinate placed at the query class, and $(1/M, \ldots, 1/M)$. The perturbation initialization does not alter these population profiles: its exponential factor is independent of the main class and has the same conditional expectation in every main class.

The class-weight concentration from Step 1 and equation 88 give, for each coordinate a and its population value p_a^0,

$$\frac{p_a}{p_a^0} \in \left[e^{-2c_0\theta} \frac{1 - c_0\theta}{1 + c_0\theta}, e^{2c_0\theta} \frac{1 + c_0\theta}{1 - c_0\theta} \right], \tag{93}$$

and the same bound holds for q. By equation 70, the relative coordinate and ℓ_1 errors in these profiles are at most $(1 - e^{-\theta})/[16(1 + e^{\theta})]$. Since

$$\frac{m_{\mathrm{KKT}}(0)}{p(0)} = (1 - p(0))(1 - e^{-\theta}) \ge \frac{1 - e^{-\theta}}{1 + e^{\theta}}, \tag{94}$$

the error is at most $m_{\mathrm{KKT}}(0)/(16p(0))$.

At population initialization,

$$G_\alpha(0) = \begin{bmatrix} p(0)^2 + (M-1)\frac{1-p(0)}{M-1}^2 & 1/M \\ 1/M & 1/M \end{bmatrix}, \qquad c_\alpha(0) = \begin{bmatrix} p(0) \\ 1/M \end{bmatrix}. \tag{95}$$

The two KKT margins at $(1,0)^\top$ are both

$$(1 - p(0))(p(0) - \frac{1 - p(0)}{M - 1}) = m_{\mathrm{KKT}}(0). \tag{96}$$

If probability vectors change by relative coordinate and ℓ_1 error at most d, then

$$|\|p\|_2^2 - \|p^0\|_2^2| \le (\|p\|_\infty + \|p^0\|_\infty)\|p - p^0\|_1 \le p(0)d(2+d), \tag{97}$$

and the same calculation gives the analogous inner-product bound. Under the preceding choice of d, the first KKT margin changes by at most $4p(0)d$ and the second by at most $8p(0)d$. Both are therefore at least $m_{\mathrm{KKT}}(0)/2$. Averaging over prompts preserves the inequalities. The β calculation is identical with the channels exchanged.

For conditioning, the correct-versus-wrong relative log-odds of p and q differ by θ at population initialization. Concentration changes this by at most $2\log((1+c_0\theta)/(1-c_0\theta))$, and the score tube changes it by at most $4c_0\theta$. Equation equation 71 leaves a difference at least $\theta/2$. Also every coordinate is bounded below by a fixed multiple of $\frac{1-p(0)}{M-1}$. Hence the mean-value theorem for $\log$ gives

$$\|p-q\|_2^2 \ge \frac{\theta^2 e^{-4c_0\theta}}{8}\left(\frac{1-c_0\theta}{1+c_0\theta}\right)^2 \frac{1-p(0)}{M-1}^2. \tag{98}$$

For probability vectors p, q,

$$\det\begin{bmatrix}\|p\|_2^2 & \langle p,q\rangle \\ \langle p,q\rangle & \|q\|_2^2\end{bmatrix} \ge \frac{1}{M}\|p-q\|_2^2. \tag{99}$$

Indeed, writing $q = p + (q-p)$ and using $\mathbf{1}^\top(q-p) = 0$, the squared norm of the component of $q-p$ orthogonal to p is at least $\|q-p\|_2^2/(M\|p\|_2^2)$; multiplying by $\|p\|_2^2$ proves equation 99. The trace is at most a fixed multiple of $1/M$ by equation 93. Combining trace, determinant, equation 98, and $\frac{1-p(0)}{M-1} \ge e^{-\theta}/M$ gives

$$\frac{\lambda_M}{C_{\mathrm{Gram}}} I_2 \preceq G_\alpha(t), G_\beta(t) \preceq 2I_2. \tag{100}$$

□

Lemma 3 (Uniform token attention inside the Phase-I tube)**.** *Suppose that, at iteration $t \le T_1$,*

$$\max_{n,\ell,h}\left|(x_\ell^{(n)})^\top\big(W_h(t) - W_h(0)\big)x_q^{(n)}\right| \le c_0\theta. \tag{101}$$

Then there exists a constant C, depending only on θ and c, such that

$$\max_{n,\ell,h}\sigma_{\ell,h}^{(n)}(t) \le \frac{C}{L}. \tag{102}$$

Proof. Consider head 1; head 2 is identical after exchanging the μ- and e-channels. From the initialization, orthogonality of the pattern families, and $|\kappa| \le c$,

$$\left|(x_\ell^{(n)})^\top W_1(0) x_q^{(n)}\right| \le \theta(1+c^2). \tag{103}$$

Together with equation 101, this gives

$$|z_{\ell,1}^{(n)}(t)| \le \theta(1+c^2) + c_0\theta =: C_0. \tag{104}$$

Hence

$$\sigma_{\ell,1}^{(n)}(t) = \frac{e^{z_{\ell,1}^{(n)}(t)}}{\sum_{r=1}^L e^{z_{r,1}^{(n)}(t)}} \le \frac{e^{C_0}}{Le^{-C_0}} = \frac{e^{2C_0}}{L}. \tag{105}$$

Since C_0 depends only on the fixed parameters θ and c, the result follows after absorbing e^{2C_0} into C. □

A.9.3 THE SCORE TUBE CANNOT BE THE FIRST FAILURE

Lemma 4 (Uniform one-step Phase-I score displacement)**.** *Suppose the uniform Phase-I concentration event holds and the score tube equation 101 holds at iteration $t < T_1$. Then, for either head and for every sampled context–query probe $(x_{\ell'}^{(m)}, x_{q'}^{(m)})$,*

$$\left|(x_{\ell'}^{(m)})^\top \Delta_t W_h x_{q'}^{(m)}\right| \le C\eta_{\mathrm{att}}\gamma^6 \left[\frac{1}{M} + c^2\left(\frac{1}{M_z} + \frac{1}{M}\right)\right]^2. \tag{106}$$

The same bound, up to a fixed numerical factor, holds for every retained centered contribution from the fifteen non-signal blocks.

Moreover, for every pair of clean main-pattern indices a, q,

$$\left|(P_\mu^{(a)})^\top \Delta_t W_1 P_\mu^{(q)}\right| \le \frac{C\eta_{\mathrm{att}}\gamma^6}{M^2}, \tag{107}$$

and the corresponding head-2 bound holds after exchanging μ and e.

Proof. We prove the head-1 statement; the head-2 argument is identical. By Lemma 1,

$$\Delta_t W_1 = \frac{\eta_{\mathrm{att}}\gamma^2}{B}\sum_{n=1}^{B}\sum_{\ell=1}^{L}\sigma_{\ell,1}^{(n)}(t)\Gamma_{\ell,1}^{(n)}(t)x_\ell^{(n)}(x_q^{(n)})^\top. \tag{108}$$

Because $\alpha, \beta \in \mathcal{V}$ and every target and retrieval vector has norm at most γ,

$$|\Gamma_{\ell,1}^{(n)}(t)| \le 8. \tag{109}$$

Lemma 3 also gives

$$\sigma_{\ell,1}^{(n)}(t) \le \frac{C}{L}. \tag{110}$$

Fix a sampled probe $(x_{\ell'}^{(m)}, x_{q'}^{(m)})$. Probing equation 108 on both sides gives

$$\begin{aligned}&\left|(x_{\ell'}^{(m)})^\top \Delta_t W_1 x_{q'}^{(m)}\right| \\ &\quad\le \frac{C\eta_{\mathrm{att}}\gamma^2}{BL}\sum_{n=1}^{B}\sum_{\ell=1}^{L}\left|(x_{\ell'}^{(m)})^\top x_\ell^{(n)}\right|\left|(x_q^{(n)})^\top x_{q'}^{(m)}\right|.\end{aligned} \tag{111}$$

By orthogonality of the four feature families, the first inner product is nonzero only when one of the corresponding feature indices agrees. After dividing by γ^2, its absolute value is bounded by a fixed multiple of

$$\mathbf{1}\{\mu_\ell = \mu_{\ell'}\} + \mathbf{1}\{e_\ell = e_{\ell'}\} + c^2\mathbf{1}\{z_{\mu,\ell} = z_{\mu,\ell'}\} + c^2\mathbf{1}\{z_{e,\ell} = z_{e,\ell'}\}. \tag{112}$$

Indeed, a perturbation match contains the product of the perturbation amplitudes of the probed and updated tokens and is therefore bounded by c^2. The same bound holds for the query inner product.

On the uniform concentration event, for every prompt and every fixed probed context index,

$$\frac{1}{L}\sum_{\ell=1}^{L}\frac{\left|(x_{\ell'}^{(m)})^\top x_\ell^{(n)}\right|}{\gamma^2} \le C\left[\frac{1}{M} + c^2\left(\frac{1}{M_z} + \frac{1}{M}\right)\right]. \tag{113}$$

Likewise, exact balance of the main query indices and the perturbation-query occupancy bound give

$$\frac{1}{B}\sum_{n=1}^{B}\frac{\left|(x_q^{(n)})^\top x_{q'}^{(m)}\right|}{\gamma^2} \le C\left[\frac{1}{M} + c^2\left(\frac{1}{M_z} + \frac{1}{M}\right)\right]. \tag{114}$$

Let

$$r := \frac{1}{M} + c^2\left(\frac{1}{M_z} + \frac{1}{M}\right).$$

Applying equation 113 inside each prompt and then equation 114 across prompts yields

$$\sum_{n=1}^{B}\sum_{\ell=1}^{L}\left|(x_{\ell'}^{(m)})^\top x_{\ell}^{(n)}\right|\left|(x_{q}^{(n)})^\top x_{q'}^{(m)}\right| \le CBL\gamma^4 r^2. \tag{115}$$

Substituting this into equation 111 proves equation 106.

We next verify that centering the non-signal main-index selectors does not change the order of the bound. For a main selector,

$$\left|\mathbf{1}\{e=e_i\}-\frac{1}{M}\right| \le \mathbf{1}\{e=e_i\}+\frac{1}{M}. \tag{116}$$

Therefore, on the same concentration event,

$$\frac{1}{L}\sum_{\ell=1}^{L}\left|\mathbf{1}\{e_\ell=e_i\}-\frac{1}{M}\right| \le \frac{C}{M}, \tag{117}$$

and the analogous batch average is also at most C/M. Perturbation selectors are already centered and satisfy

$$|\kappa\mathbf{1}\{z=z_k\}| \le c\,\mathbf{1}\{z=z_k\}.$$

Thus replacing any main selector by its centered version changes only the absolute constant in equation 113–equation 114. Since the leakage decomposition contains only a fixed number of retained terms from the fifteen non-signal blocks, the same bound holds for each centered leakage contribution and for their sum after adjusting C.

Finally, probe equation 108 by $P_\mu^{(a)}$ and $P_\mu^{(q)}$. Orthogonality gives

$$\frac{|\langle P_\mu^{(a)}, x_\ell^{(n)}\rangle|}{\gamma^2} = \mathbf{1}\{\mu_\ell^{(n)}=\mu_a\},$$

and similarly on the query side. The context occupancy is at most CL/M and exact query balance gives exactly B/M prompts with query class q. Therefore

$$\left|(P_\mu^{(a)})^\top \Delta_t W_1 P_\mu^{(q)}\right| \le \frac{C\eta_{\mathrm{att}}\gamma^6}{M^2}, \tag{118}$$

which proves equation 107. □

Corollary 2 (Persistence of the Phase-I score tube)**.** *Suppose*

$$C\eta_{\mathrm{att}}\gamma^6 T_1\left[\frac{1}{M}+c^2\left(\frac{1}{M_z}+\frac{1}{M}\right)\right]^2 \le c_0\theta. \tag{119}$$

Then, on the Phase-I concentration event,

$$\max_{0\le t\le T_1}\max_{n,\ell,h}\left|(x_\ell^{(n)})^\top\big(W_h(t)-W_h(0)\big)x_q^{(n)}\right| \le c_0\theta. \tag{120}$$

Moreover, at $t=T_1$,

$$|A_{s,q}(T_1)-A_{r,q}(T_1)| \le \frac{c_{\mathrm{wrong}}}{4}, \qquad r,s\ne q, \tag{121}$$

and

$$A_{q,q}(T_1)-A_{s,q}(T_1) \ge \theta - 2c_0\theta, \qquad s\ne q, \tag{122}$$

provided the numerical constant in equation 119 is chosen to satisfy the corresponding c_{wrong} budget.

Proof. Let τ be the first iteration at which equation 120 fails. If $\tau \le T_1$, then the tube holds at every $t<\tau$, so Lemma 4 applies at all such iterations. Since $W_h(0)-W_h(0)=0$,

$$\left|(x_\ell^{(n)})^\top\big(W_h(\tau)-W_h(0)\big)x_q^{(n)}\right| \le \sum_{t=0}^{\tau-1}\left|(x_\ell^{(n)})^\top\Delta_t W_h x_q^{(n)}\right| \tag{123}$$

$$\le C\eta_{\mathrm{att}}\gamma^6 T_1\left[\frac{1}{M}+c^2\left(\frac{1}{M_z}+\frac{1}{M}\right)\right]^2 \tag{124}$$

$$\le c_0\theta, \tag{125}$$

contradicting the definition of τ. Thus the score tube persists through T_1.

For the clean block, Lemma 4 gives

$$|A_{a,q}(T_1) - A_{a,q}(0)| \le \frac{C\eta_{\mathrm{att}}\gamma^6 T_1}{M^2}. \tag{126}$$

Since

$$A_{q,q}(0) = \theta, \qquad A_{s,q}(0) = 0, \quad s \neq q,$$

the assumed Phase-I step-size budget gives equation 121 and equation 122. □

A.9.4 Contraction and exact clipping of the value coefficients

Lemma 5 (Value identification)**.** *On the Phase-I concentration event, if the score tube holds through T_1, then*

$$\alpha(T_1) = (1,0)^\top, \qquad \beta(T_1) = (0,1)^\top.$$

Proof. By equation 48, the unconstrained value gradient is $\gamma^2(G_\alpha(t)\alpha(t) - c_\alpha(t))$, with the corresponding channel-exchanged formula for β. Put $\alpha^\star = (1,0)^\top$. The positive KKT margins give the projected-gradient fixed-point identity

$$\alpha^\star = \mathrm{Proj}_{\mathcal{V}}\left(\alpha^\star - \eta_v\gamma^2(G_\alpha(t)\alpha^\star - c_\alpha(t))\right). \tag{127}$$

Projection is non-expansive. Since equation 100 holds and $\eta_v\gamma^2 \le 1/2$,

$$\|\alpha(t+1) - \alpha^\star\|_2 \le \left(1 - \frac{\eta_v\gamma^2\lambda_M}{C_{\mathrm{Gram}}}\right)\|\alpha(t) - \alpha^\star\|_2. \tag{128}$$

Thus equation 72 gives

$$\|\alpha(T_{\mathrm{pre}}) - \alpha^\star\|_2 \le \frac{m_{\mathrm{KKT}}(0)\eta_v\gamma^2}{32}. \tag{129}$$

Let $x = \alpha(T_{\mathrm{pre}})$ and let g be the value gradient on the next fresh batch. At $\alpha^\star$, the KKT margins give

$$g_1(\alpha^\star) \le -\frac{\gamma^2 m_{\mathrm{KKT}}(0)}{2}, \qquad g_2(\alpha^\star) - g_1(\alpha^\star) \ge \frac{\gamma^2 m_{\mathrm{KKT}}(0)}{2}. \tag{130}$$

Since $G_\alpha \preceq 2I_2$, equation 129 implies

$$g_1(x) \le -\frac{\gamma^2 m_{\mathrm{KKT}}(0)}{4}, \qquad g_2(x) - g_1(x) \ge \frac{\gamma^2 m_{\mathrm{KKT}}(0)}{4}. \tag{131}$$

For $y = x - \eta_v g(x)$, equations equation 129–equation 131 give $y_1 \ge 1$ and $y_1 - y_2 \ge 1$. These are exactly the two normal-cone inequalities for $\Pi_{\mathcal{V}}(y) = (1,0)^\top$: the feasible cone at $(1,0)^\top$ is generated by $(-1,0)^\top$ and $(-1,1)^\top$. Therefore

$$\alpha(T_1) = (1,0)^\top. \tag{132}$$

The channel-exchanged proof gives $\beta(T_1) = (0,1)^\top$. Since $m_{\mathrm{KKT}}(0) = \Theta_\theta(M^{-1})$ and $\lambda_M = \Theta_\theta(M^{-2})$, equation 72 gives the stated bound on T_1. □

A.9.5 Cross-query synchronization at the handoff

By equation (258), simultaneously for all $q \neq q'$,

$$\max_{t \le T_1} |\text{one probed coefficient} - \text{its } q \leftrightarrow q' \text{ relabeling}| \le \frac{c_{\mathrm{sync}}}{C_{\mathrm{hand}}M}. \tag{133}$$

For $s \neq q$, set

$$\Delta_{s,q}(T_1) := A_{s,q}(T_1) - A_{q,q}(T_1).$$

Let τ be the transposition exchanging q and q'. Since each $\Delta_{s,q}$ is the difference of two probed coefficients, equation 258 gives

$$\left|\Delta_{s,q}(T_1) - \Delta_{\tau(s),q'}(T_1)\right| \le \frac{2c_{\mathrm{sync}}}{C_{\mathrm{hand}}M}. \tag{134}$$

Moreover, the clean preferred-gap estimate gives

$$\Delta_{s,q}(T_1) \le -\theta + 2c_0\theta, \qquad s \neq q. \tag{135}$$

Using

$$\frac{1-p_q(T_1)}{p_q(T_1)} = \sum_{s\neq q} e^{\Delta_{s,q}(T_1)} \tag{136}$$

and relabeling the q' sum by τ, the mean-value theorem gives

$$\begin{aligned}&\left|\frac{1-p_q(T_1)}{p_q(T_1)} - \frac{1-p_{q'}(T_1)}{p_{q'}(T_1)}\right| \\ &\quad\le \frac{2(M-1)c_{\mathrm{sync}}}{C_{\mathrm{hand}}M}\exp\left(-\theta + 2c_0\theta + \frac{2c_{\mathrm{sync}}}{C_{\mathrm{hand}}M}\right).\end{aligned} \tag{137}$$

Choosing the fixed constant C_{hand} sufficiently large therefore gives

$$\max_{q,q'}\left|\frac{1-p_q(T_1)}{p_q(T_1)} - \frac{1-p_{q'}(T_1)}{p_{q'}(T_1)}\right| \le \frac{c_{\mathrm{sync}}}{16}. \tag{138}$$

The head-2 statement follows by exchanging the μ- and e-channels.

A.9.6 Leakage and preferred-class maximality at the handoff

Apply the componentwise form of equation 106 to each of the fifteen non-signal blocks and center every main-index selector by subtracting its uniform mean; perturbation selectors are already centered because $\mathbb{E}\kappa = 0$. Centering changes only the numerical constant. Since the learned deviation is zero at time zero, equation 74 gives

$$\max|\text{centered non-signal score contributed by } W_h(T_1) - W_h(0)| \le \frac{c_{\mathrm{leak}}}{8C_{\mathrm{prod}}}. \tag{139}$$

To convert this score bound into a class-normalizer bound, note that for any collection of clean scores $\{z_\ell\}_{\ell\in C}$ and perturbations satisfying $|\delta_\ell| \le r$,

$$e^{-r}\sum_{\ell\in C} e^{z_\ell} \le \sum_{\ell\in C} e^{z_\ell+\delta_\ell} \le e^{r}\sum_{\ell\in C} e^{z_\ell}. \tag{140}$$

Taking logarithms gives

$$\left|\log\sum_{\ell\in C} e^{z_\ell+\delta_\ell} - \log\sum_{\ell\in C} e^{z_\ell}\right| \le r. \tag{141}$$

Applying equation 141 with

$$r = \frac{c_{\mathrm{leak}}}{8C_{\mathrm{prod}}}$$

shows that the learned leakage changes each class log-normalizer by at most $c_{\mathrm{leak}}/(8C_{\mathrm{prod}})$. Hence the difference between the learned leakage corrections of any two classes is at most

$$\frac{c_{\mathrm{leak}}}{4C_{\mathrm{prod}}}. \tag{142}$$

It remains to account for the fixed perturbation initialization. Conditional on the query perturbation, its exponential factor has the same distribution in every main class and is independent of the main index. Therefore the second concentration estimate in Step 1 and equation 84 imply that the difference of its corrections to any two class log-normalizers is at most

$$\frac{c_{\mathrm{leak}}}{8C_{\mathrm{prod}}}.$$

Combining the learned and fixed contributions gives

$$\frac{c_{\mathrm{leak}}}{4C_{\mathrm{prod}}} + \frac{c_{\mathrm{leak}}}{8C_{\mathrm{prod}}} = \frac{3c_{\mathrm{leak}}}{8C_{\mathrm{prod}}} < \frac{c_{\mathrm{leak}}}{2C_{\mathrm{prod}}}. \tag{143}$$

Thus the total class-dependent non-signal correction to any preferred-versus-wrong class log-mass difference is at most $c_{\mathrm{leak}}/(2C_{\mathrm{prod}})$.

For a head-1 prompt with μ-query class q and any $s \neq q$, the full class log-mass difference equals the clean difference $A_{q,q} - A_{s,q}$ plus the class-dependent non-signal correction. By equation 121, the leakage bound just proved, and equation 76,

$$\log \frac{\pi_q^{(n)}(T_1)}{\pi_s^{(n)}(T_1)} \geq \theta - 2c_0\theta - \frac{c_{\mathrm{leak}}}{2C_{\mathrm{prod}}} \geq \frac{\theta}{2} > 0. \tag{144}$$

Thus the preferred class is strictly maximal. Exchanging μ and e proves the head-2 statement. All six conclusions of Proposition 2 now hold on the single event from Step 1. □

Phase-I output. Proposition 2 supplies the base case for the Phase-II induction. At $t = T_1$, the values are exactly specialized, the preferred class is maximal, wrong clean scores are controlled, cross-query differences are synchronized, the KKT margins are positive, and all centered non-signal scores lie strictly inside the leakage tube. Phase II shows that these properties persist while the clean preferred attention grows to the target level.

A.10 Phase II: finite-sample training dynamics

Scope. For a fixed horizon T, construct the T training batches independently. The finite training set contains

$$N = BT \tag{145}$$

prompts. Before optimization they are organized into T disjoint batches, each generated independently by the balanced query design:

$$\mathcal{B}_0, \ldots, \mathcal{B}_{T-1}, \qquad |\mathcal{B}_t| = B, \tag{146}$$

Batch $\mathcal{B}_t$ is used once to produce the update from iteration t to iteration $t+1$. In addition, sample one independent terminal evaluation batch $\mathcal{B}_T$ from the same balanced query design. Conditional expectations at iteration t condition on the initialization and on $\mathcal{B}_0, \ldots, \mathcal{B}_{t-1}$, so the permutation blocks and all remaining variables in $\mathcal{B}_t$ are fresh. The terminal batch is not used for optimization. The proof is written for head 1; head 2 follows by exchanging the μ- and e-channels.

A.11 Primary Phase-II quantities

We retain $A(t)$, $p_q(t)$, and $a_q(t)$ from equation 89–equation 31. We use $a_q(t)$ as the dynamical log-odds state; full-prompt odds are written directly as $p_n(t)/(1-p_n(t))$ when they are needed.

For a prompt n at time t, let j_n be its μ-query class, let $\pi_a^{(n)}(t)$ be its full head-1 attention mass on μ-class a, and put

$$p_n(t) := \pi_{j_n}^{(n)}(t), \qquad d_n(t) := p_n(t)(1-p_n(t))^2. \tag{147}$$

After value specialization, differentiating the prompt loss with respect to the class logits gives

$$h_a^{(n)}(t) = \pi_a^{(n)}(t)\left(\mathbf{1}\{a = j_n\} - p_n(t) - \pi_a^{(n)}(t) + \sum_{b=1}^{M}(\pi_b^{(n)}(t))^2\right). \tag{148}$$

Under preferred-class maximality, the exact identities and inequalities used below are

$$\sum_a h_a^{(n)} = 0, \quad h_s^{(n)} \leq 0 \ (s \neq j_n), \quad d_n \leq h_{j_n}^{(n)} \leq 2d_n, \quad |h_a^{(n)}| \leq 4d_n. \tag{149}$$

For $s \neq q$, set $S_{q,s} = A_{q,q} - A_{s,q}$. Exact batch balance gives

$$S_{q,s}(t+1) - S_{q,s}(t) = \frac{\eta_{\mathrm{att}}\gamma^6}{B} \sum_{n:j_n=q} \left(h_q^{(n)} - h_s^{(n)}\right) \geq \frac{\eta_{\mathrm{att}}\gamma^6}{B} \sum_{n:j_n=q} d_n(t). \tag{150}$$

THE FIFTEEN CENTERED BLOCKS

For head 1, arrange the attention matrix into the context/query feature groups μ, e, z_μ, z_e. The μ–μ block is the clean signal block. The other fifteen blocks are

group	number	blocks
query-side	3	μ–e, μ–z_μ, μ–z_e
context-side	3	e–μ, z_μ–μ, z_e–μ
two-sided	9	u–v, $u, v \in \{e, z_\mu, z_e\}$

For an e index use the centered selector $\mathbf{1}\{e = e_i\} - 1/M$. For a perturbation index use $\kappa\mathbf{1}\{z = z_k\}$, which is already centered because $\mathbb{E}\kappa = 0$. A centered leakage score is the contribution of one of these fifteen blocks after these centerings. This is the centered quantity used in the Phase-I handoff. The raw fixed perturbation score in $W_1(0)$ is not included; its class-normalizer correction is controlled on the all-horizon concentration event of Lemma 6. More explicitly, decompose every nuisance main selector as

$$\mathbf{1}\{e = e_i\} = \left(\mathbf{1}\{e = e_i\} - \frac{1}{M}\right) + \frac{1}{M}. \tag{151}$$

In a block with two nuisance main sides, this gives the centered–centered term and the centered-context–mean-query term. The mean-context–centered-query and mean–mean terms are constant across context tokens and hence they do not affect attention and their gradients are zero. We retain both nongauge terms in the leakage vector. Perturbation selectors have zero mean and are evaluated with their factors $|\kappa| \le c$. The leakage maximum norm is the supremum of the sum of all retained centered contributions over the allowed token and query indices and amplitudes. Thus it controls the actual non-signal difference between any two context-token scores. This centering only splits a raw block into a fixed number of coordinates, so the total coordinate count remains $C(M + M_z)^2$. Head 2 uses the same definition after exchanging μ and e.

A.12 HANDOFF CONDITIONS

Invoke the existing Phase-I handoff with failure probability $\delta/4$. At $t = T_1$ it gives value specialization, positive KKT margins, the $c_{\mathrm{wrong}}/4$ wrong-score tube, preferred-class maximality with full log-mass gap at least $\theta/2$, inverse-odds synchronization, and the stated centered leakage budget. We also use equation equation 258 from the proof of that same proposition: every corresponding probed coefficient under a $q \leftrightarrow q'$ exchange is within $c_{\mathrm{sync}}/(C_{\mathrm{hand}}M)$ at T_1. We retain this intermediate estimate because it initializes the early Phase-II exchange argument. We choose the constants in equation 133 such that each following inequality holds with strict margin.

Probabilities. Proposition 2 is invoked with failure probability $\delta/4$. In Phase II we allocate $\delta/8$ to the all-horizon clean-to-prompt event of Lemma 6, $\delta/8$ to the leakage event of Lemma 9, and $\delta/4$ to synchronization. The synchronization budget contains three events of size at most $\delta/16$: the early exchanged-coordinate bound, the early inverse-odds martingale, and the late martingale excursion. Thus the total Phase-II failure probability is at most $\delta/2$, and the handoff and Phase-II events together fail with probability at most $3\delta/4 < \delta$.

Stopped Phase-II interval. Let τ be the first iteration $t \ge T_1$ at which any of the following fails: the values are specialized; the preferred full class is maximal; the clean wrong-class differences lie in the c_{wrong} tube; every centered leakage score is smaller than c_{leak}; the appropriate synchronization bound equation 172–equation 173 holds; or the KKT conditions at the specialized value vertices remain nonnegative. Set

$$\tau := \min\{\text{first such failure}, T + 1\}.$$

By Proposition 2, all these conditions hold with strict slack at $t = T_1$. All Phase-II estimates below are proved on $T_1 \le t < \tau$. Proposition 3 shows that no boundary can be the first one reached, and hence $\tau > T$.

A.13 CLEAN-TO-PROMPT COMPARISON

Lemma 6 (Clean reference versus actual prompt mass)**.** *On the stopped interval $T_1 \le t < \tau$, if,*

$$L \ge CM \log \frac{C(M+M_z)^2 B(T+1)}{\delta}, \tag{152}$$

then, with probability at least $1-\delta/8$, simultaneously for every used prompt, every $t \le T$, and $q = j_n$,

$$\left| \log \frac{p_n(t)/(1-p_n(t))}{p_q(t)/(1-p_q(t))} \right| \le \rho. \tag{153}$$

Here ρ is a fixed constant made sufficiently small by the constants in equation 152 and equation 33. Consequently

$$e^{-2\rho} p_q(t)(1-p_q(t))^2 \le d_n(t) \le e^{2\rho} p_q(t)(1-p_q(t))^2. \tag{154}$$

On the same event, uniformly over every prompt in $\mathcal{B}_0, \dots, \mathcal{B}_T$ and every pair of main classes a, b, after factoring out the clean factors $e^{A_{a,q}(t)}$ and $e^{A_{b,q}(t)}$, the learned centered contribution to the difference of the two class log-normalizers is at most $2c_{\mathrm{leak}}$. Moreover, the part of this difference due to the fixed perturbation initialization is at most

$$\frac{c_{\mathrm{leak}}}{8}. \tag{155}$$

Consequently the non-clean difference is at most

$$2c_{\mathrm{leak}} + \frac{c_{\mathrm{leak}}}{8} < 3c_{\mathrm{leak}}. \tag{156}$$

Proof. Condition on $\mathcal{F}_t$, the history before the fresh batch. In every main class a, factor $e^{A_{a,q}(t)}$ from the class partition sum. The remainder is an average of independent positive nuisance weights. The fixed initialization, stopped leakage bound, and $|\kappa| \le c$ give fixed bounds $0 < c_1 \le Z \le c_2$. The mean is the same in every main class because main indices, nuisance indices, and amplitudes are mutually independent.

Bernstein's inequality and the binomial class-count bound give, simultaneously,

$$\left| \frac{\widehat{Z}_a}{\mathbb{E}_t Z_a} - 1 \right| \le C \left(\sqrt{\frac{Mu}{L}} + \frac{Mu}{L} \right), \qquad u = \log \frac{C(M+M_z)^2 B(T+1)}{\delta}. \tag{157}$$

A union bound covers both heads, all classes, prompts, and iterations. Query-side leakage changes a class log-weight by at most c_{leak}. Context-side and two-sided terms stay inside the bounded sample means.

The fixed perturbation initialization is controlled on the same all-T-batch event, not by the Phase-I handoff event. Conditional on the query perturbation, its exponential factor is independent of the main class and has the same conditional expectation in every main class. Hence the same multiplicative Bernstein–Chernoff estimate used above applies simultaneously to both heads, all B(T+1) prompts in the T training batches together with the terminal evaluation batch, and all main classes.

With

$$u = \log \frac{C(M+M_z)^2 B(T+1)}{\delta}, \tag{158}$$

it gives

$$\left| \frac{\widehat{Z}_a}{\mathbb{E}_t Z_a} - 1 \right| \le C \left(\sqrt{\frac{Mu}{L}} + \frac{Mu}{L} \right) \tag{159}$$

for every such class weight. The numerical constant in equation 152 is chosen large enough that

$$2 \log \frac{1 + C\left(\sqrt{Mu/L} + Mu/L\right)}{1 - C\left(\sqrt{Mu/L} + Mu/L\right)} \le \frac{c_{\mathrm{leak}}}{8}. \tag{160}$$

The current learned centered leakage changes any class log-normalizer by at most c_{leak}, so its contribution to the difference of two class log-normalizers is at most $2c_{\mathrm{leak}}$. Combining this with the

fixed-initialization bound equation 155 gives the bound equation 156. The union bound in equation 152 contains the factor $B(T+1)$, so this conclusion holds simultaneously for all training batches and the terminal evaluation batch.

Hence the actual preferred-versus-aggregate-wrong log-odds differs from equation 31 by

$$C\left(c_{\mathrm{leak}}+\sqrt{\frac{Mu}{L}}+\frac{Mu}{L}\right), \tag{161}$$

which proves equation 153. If two log-odds differ by at most ρ, the ratios of their probabilities and complementary probabilities belong to $[e^{-\rho}, e^{\rho}]$. This proves equation 154. □

A.14 Clean attention progress

For $s \neq q$, let

$$\omega_{s,q}(t)=\frac{e^{A_{s,q}(t)}}{\sum_{r\neq q} e^{A_{r,q}(t)}}. \tag{162}$$

Lemma 7 (Exact average clean progress)**.** *Before the joint stopping time,*

$$a_q(t+1)-a_q(t)=-\log\sum_{s\neq q}\omega_{s,q}(t)e^{-[S_{q,s}(t+1)-S_{q,s}(t)]}. \tag{163}$$

$$a_q(t+1)-a_q(t)\geq\frac{\eta_{\mathrm{att}}\gamma^6}{B}\sum_{n:j_n=q} d_n(t). \tag{164}$$

Therefore

$$\boxed{\frac{1}{M}\sum_{q=1}^{M}\big(a_q(t+1)-a_q(t)\big)\geq\frac{\eta_{\mathrm{att}}\gamma^6}{M}\left[\frac{1}{B}\sum_{n=1}^{B} d_n(t)\right].} \tag{165}$$

Moreover, for sufficiently small $\eta_{\mathrm{att}}\gamma^6$,

$$0\leq a_q(t+1)-a_q(t)\leq\frac{C\eta_{\mathrm{att}}\gamma^6}{B}\sum_{n:j_n=q} d_n(t)\leq 1. \tag{166}$$

Proof. Subtract equation 31 at consecutive times. Write every wrong-score increment as the preferred-score increment minus the corresponding clean-gap increment. The common preferred increment cancels and gives equation 163. Every gap increment is at least the right side of equation 164 by equation 150, and the $\omega_{s,q}$ sum to one. This proves equation 164; summing over q proves equation 165. Finally, $h_q-h_s\leq|h_q|+|h_s|\leq 6d_n$ by equation 149, and equation 163 lies between the smallest and largest gap increments. □

Lemma 8 (Finite squared-signal energy)**.** *If the class signals are synchronized up to a fixed factor and*

$$G(t):=\max_q\frac{M}{B}\sum_{n:j_n=q} d_n(t), \tag{167}$$

then on every stopped Phase-II interval

$$\sum_t\left(\frac{\eta_{\mathrm{att}}\gamma^6}{M}G(t)\right)^2\leq\frac{C\eta_{\mathrm{att}}\gamma^6}{M}. \tag{168}$$

Proof. Synchronization and Lemma 6 make $G(t)$ at most a fixed multiple of $p_q(t)(1-p_q(t))^2$ for every q. Lemma 7 bounds $(\eta_{\mathrm{att}}\gamma^6/M)G(t)$ by a fixed multiple of $a_q(t+1)-a_q(t)$. The one-step bound equation 166 controls within-step variation, so

$$\begin{aligned}\sum_t\left(\frac{\eta_{\mathrm{att}}\gamma^6}{M}G(t)\right)^2&\leq\frac{C\eta_{\mathrm{att}}\gamma^6}{M}\sum_t p_q(t)(1-p_q(t))^2[a_q(t+1)-a_q(t)]\\&\leq\frac{C\eta_{\mathrm{att}}\gamma^6}{M}\int_{-\infty}^{\infty}p(a)(1-p(a))^2\,da.\end{aligned} \tag{169}$$

Since $dp/da = p(1-p)$, the integral is $\int_0^1 (1-p)\,dp = 1/2$. On the event of Lemma 6, exact query balance also gives

$$\max_q p_q(t)(1-p_q(t))^2 \le CG(t). \tag{170}$$

The left side is $\mathcal{F}_t$-measurable. Hence the conditional finite-context and fresh-batch variance bounds below may be expressed in terms of this predictable quantity, Lemma 8, and equation 170 then control its cumulative energy. □

A.15 Control of non-signal attention scores

Let $\mathcal{F}_t$ contain initialization and all batches before iteration t.

Lemma 9 (Master max-norm leakage bound)**.** *On the stopped interval $T_1 \le t < \tau$, assume the Phase-I handoff bound*

$$\max |\textit{centered leakage score at } T_1| \le \frac{c_{\mathrm{leak}}}{4C_{\mathrm{prod}}}.$$

If

$$L, B \ge CM \log \frac{C(M+M_z)^2 B(T+1)}{\delta}, \qquad M_z \ge Cc^2 M, \tag{171}$$

and $\eta_{\mathrm{att}}\gamma^6$ is sufficiently small, then, with probability at least $1-\delta/8$, simultaneously for $T_1 \le t \le T$, every centered leakage score of both heads is strictly smaller than c_{leak}.

Proof. Centering removes all class-independent softmax gauge terms. Appendix B shows that the context-side and two-sided leakage blocks have zero infinite-context drift. For the three query-side blocks Q_e, Q_{z_μ}, and Q_{z_e}, the infinite-context updates are nonexpansive once every preferred clean mass is at least $1/2$. Before that time, their one-step amplification is at most

$$1 + \frac{C\eta_{\mathrm{att}}\gamma^6}{M^2} \max_q p_q(t)(1-p_q(t))^2,$$

whose product over time is bounded by the fixed constant C_{prod}.

Appendix B also proves that the accumulated finite-context contribution to every centered leakage score is at most $c_{\mathrm{leak}}/8$, and that, with probability at least $1-\delta/8$, the accumulated centered fresh-batch contribution is at most $c_{\mathrm{leak}}/8$ uniformly over the training horizon. The Phase-I handoff contributes at most $c_{\mathrm{leak}}/4$ after the possible population amplification. Hence, for every $T_1 \le t \le T$,

$$\max |\text{centered leakage score at time } t| \le \frac{c_{\mathrm{leak}}}{4} + \frac{c_{\mathrm{leak}}}{8} + \frac{c_{\mathrm{leak}}}{8} < c_{\mathrm{leak}}.$$

Thus the leakage boundary cannot be reached before time T. □

A.16 Synchronization across query classes

Synchronization uses two regimes. Before the preferred masses reach $1/2$, we couple the centered clean wrong-class shape and the three query-side leakage blocks under class exchange. After the crossing, the clean log-odds drift itself provides the restoring force.

Lemma 10 (Synchronization persistence)**.** *On the stopped interval $T_1 \le t < \tau$, under equation 152, and sufficiently large constants, using the bounds in equation 33 and in Proposition 2, with probability at least $1-\delta/4$,*

$$\max_{q,q'} \left| \frac{1-p_q(t)}{p_q(t)} - \frac{1-p_{q'}(t)}{p_{q'}(t)} \right| \le \frac{c_{\mathrm{sync}}}{2}, \qquad \min_q p_q(t) \le \frac{1}{2}, \tag{172}$$

and

$$\max_{q,q'} |a_q(t) - a_{q'}(t)| \le c_{\mathrm{sync}}, \qquad \min_q p_q(t) \ge \frac{1}{2}. \tag{173}$$

Hence the class signals are comparable by a fixed factor throughout the stopped interval.

Proof. The proof has two regimes.

When some preferred clean mass is below $1/2$, synchronization is measured in inverse odds. Under a $q \leftrightarrow q'$ exchange, only the centered clean wrong-class shape and the three query-side centered blocks require an exchange estimate. Their combined exchanged effect remains $O(c_{\mathrm{sync}}/M)$ up to the $1/2$ crossing. The clean inverse-odds map is nonexpansive, the finite-context correction is Lipschitz in the pair difference, and the remaining batch fluctuation is controlled uniformly in time. Hence

$$\max_{q,q'} \left| \frac{1-p_q(t)}{p_q(t)} - \frac{1-p_{q'}(t)}{p_{q'}(t)} \right| \le \frac{c_{\mathrm{sync}}}{2} \qquad \text{while } \min_q p_q(t) \le \frac{1}{2}. \tag{174}$$

Once every preferred clean mass is at least $1/2$, no coefficientwise exchange estimate is needed. The centered query-side selectors have zero mean, so the first-order leakage correction to the population drift cancels and the remaining effect is quadratic in the leakage radius. Meanwhile

$$\frac{d}{da}\big[p(1-p)^2\big] = p(1-p)^2(1-3p) \le -\frac{1}{2}p(1-p)^2, \qquad p \ge \frac{1}{2}, \tag{175}$$

which gives a restoring drift for the clean log-odds difference on the outer half of the synchronization bound. Finite-context corrections and the batch martingale are smaller than this restoring drift under the stated L, B conditions. Therefore

$$\max_{q,q'} |a_q(t) - a_{q'}(t)| \le c_{\mathrm{sync}} \qquad \text{while } \min_q p_q(t) \ge \frac{1}{2}. \tag{176}$$

The early exchange calculation, finite-context interpolation, and late martingale excursion argument are given in Appendix A.22. □

Proposition 3 (Closed joint Phase-II induction)**.** *On the intersection of the Phase-I handoff event and the Phase-II events of Lemmas 6, 9, and 10, for every $T_1 \le t \le T$: values remain specialized, every preferred full class remains maximal, wrong clean scores stay in the c_{wrong} tube, all fifteen centered blocks stay in the c_{leak} tube, and synchronization equation 172–equation 173 persists.*

Proof. Let τ be the stopping time defined in Section A.12. By construction, all bootstrap conditions hold on $[T_1, \tau)$, so Lemmas 6, 9, and 10 apply throughout this interval.

For wrong classes $r, s \ne j_n$, direct subtraction in equation 148 gives

$$h_r^{(n)} - h_s^{(n)} = -(\pi_r^{(n)} - \pi_s^{(n)}) \left(p_n - \sum_a (\pi_a^{(n)})^2 + \pi_r^{(n)} + \pi_s^{(n)} \right). \tag{177}$$

The parenthesis is nonnegative under preferred maximality, and the exact clean wrong-score update is

$$\begin{aligned} &[A_{r,q}(t+1) - A_{s,q}(t+1)] - [A_{r,q}(t) - A_{s,q}(t)] \\ &\quad = \frac{\eta_{\mathrm{att}}\gamma^6}{B} \sum_{n: j_n = q} (h_r^{(n)} - h_s^{(n)}). \end{aligned} \tag{178}$$

Suppose first that

$$A_{r,q}(t) - A_{s,q}(t) \ge c_{\mathrm{wrong}} - 4c_{\mathrm{leak}}. \tag{179}$$

By equation 156, the full wrong-class log-mass difference differs from this clean difference by less than $3c_{\mathrm{leak}}$. Hence

$$\log \frac{\pi_r^{(n)}(t)}{\pi_s^{(n)}(t)} \ge c_{\mathrm{wrong}} - 7c_{\mathrm{leak}} > 0, \tag{180}$$

because $c_{\mathrm{wrong}} > 8c_{\mathrm{leak}}$. Therefore $\pi_r^{(n)}(t) \ge \pi_s^{(n)}(t)$ for every prompt in the q stratum, and equation 178 gives

$$A_{r,q}(t+1) - A_{s,q}(t+1) \le A_{r,q}(t) - A_{s,q}(t). \tag{181}$$

The same argument with r and s exchanged shows that the update points upward whenever

$$A_{r,q}(t) - A_{s,q}(t) \le -c_{\mathrm{wrong}} + 4c_{\mathrm{leak}}. \tag{182}$$

It remains to rule out a jump across the boundary from the interior. By equation 149, exact query balance, and $d_n \le 1$,

$$|[A_{r,q}(t+1) - A_{s,q}(t+1)] - [A_{r,q}(t) - A_{s,q}(t)]| \le \frac{C\eta_{\mathrm{att}}\gamma^6}{M}. \tag{183}$$

Choose the already-small attention step size so that

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{M} \le c_{\mathrm{leak}}. \tag{184}$$

If the clean wrong-score difference lies in

$$(-c_{\mathrm{wrong}} + 4c_{\mathrm{leak}},\ c_{\mathrm{wrong}} - 4c_{\mathrm{leak}}), \tag{185}$$

then equation 183–equation 184 prevent it from crossing either boundary $\pm c_{\mathrm{wrong}}$ in one step. If it lies in either boundary strip, the preceding sign calculation makes its update point toward the interior. Therefore the c_{wrong} wrong-score tube cannot be the first failure.

Every clean preferred-versus-wrong gap is non-decreasing by equation 150.

By the all-T-batch event in Lemma 6, every full class-log-mass difference differs from the corresponding clean class-logit difference by at most $3c_{\mathrm{leak}}$, and the fixed-initialization part of this correction is itself bounded by equation 155. The resulting bound is therefore uniform over the full Phase-II horizon. Equation equation 33 therefore leaves a strictly positive full preferred-versus-wrong gap, so preferred maximality cannot be the first failure.

At the value vertex, the first head-1 KKT margin is

$$p_n - \sum_a (\pi_a^{(n)})^2 \ge 0 \tag{186}$$

under preferred maximality. For the cross-head margin, let π be head 1's μ profile and $\widetilde{\pi}$ head 2's μ profile. Its normalized value is

$$(e_{j_n} - \pi)^\top (\pi - \widetilde{\pi}). \tag{187}$$

The head-2 μ profile is a small multiplicative perturbation of the uniform profile, while head 1 has a positive preferred gap and comparable wrong entries. For exactly symmetric profiles this display equals

$$(1 - p_n)\left(p_n - \frac{1 - p_n}{M - 1}\right) > 0. \tag{188}$$

Using

$$\bar{r} = \frac{1 - p_n}{M - 1}, \tag{189}$$

the score tubes give

$$|\pi_s - \bar{r}| \le C(c_{\mathrm{wrong}} + c_{\mathrm{leak}} + \rho)\bar{r}, \qquad s \ne j_n, \tag{190}$$

and

$$|\widetilde{\pi}_s - 1/M| \le \frac{C(c_{\mathrm{leak}} + \rho)}{M}. \tag{191}$$

Expanding the inner product and using that both profiles sum to one gives

$$\left|(e_{j_n} - \pi)^\top (\pi - \widetilde{\pi}) - (1 - p_n)(p_n - \bar{r})\right| \le C(c_{\mathrm{wrong}} + c_{\mathrm{leak}} + \rho)(1 - p_n)(p_n - \bar{r}). \tag{192}$$

The numerical choices in equation 33 and equation 152 make the coefficient on the right strictly smaller than one. Hence both head-1 KKT margins are nonnegative. Exchanging the μ and e channels proves the head-2 margins. Projected gradient descent therefore fixes both specialized vertices.

Lemma 9 excludes a first failure of the leakage bound, and Lemma 10 excludes a first failure of either synchronization condition. The preceding deterministic arguments exclude a first failure of the wrong-score tube, preferred maximality, or KKT persistence. Therefore no boundary in the definition of τ can be reached, and $\tau > T$. □

A.17 Proof of Theorem 1

Set

$$\overline{a}(t) := \frac{1}{M}\sum_{q=1}^{M} a_q(t), \qquad \overline{p}(t) := \frac{e^{\overline{a}(t)}}{1+e^{\overline{a}(t)}}. \tag{193}$$

Synchronization and Lemma 6 give

$$\frac{1}{B}\sum_{n=1}^{B} d_n(t) \geq \frac{1}{C}\,\overline{p}(t)(1-\overline{p}(t))^2. \tag{194}$$

Combining equation 165 and equation 194,

$$\overline{a}(t+1) - \overline{a}(t) \geq \frac{\eta_{\mathrm{att}}\gamma^6}{CM}\overline{p}(t)(1-\overline{p}(t))^2. \tag{195}$$

A.17.1 Training Convergence

The Phase-I handoff gives

$$\frac{1-p_q(T_1)}{p_q(T_1)} \leq (M-1)e^{-(1-2c_0)\theta}, \tag{196}$$

and synchronization gives the same order for $e^{-\overline{a}(T_1)}$.

First suppose $\overline{p}(t) \leq 1/2$ and, only in this calculation, put $s(t) = e^{-\overline{a}(t)} \geq 1$. The one-step increment is at most one, so

$$\begin{aligned} s(t) - s(t+1) &= s(t)\left(1 - e^{-[\overline{a}(t+1)-\overline{a}(t)]}\right) \\ &\geq \frac{s(t)}{2}[\overline{a}(t+1) - \overline{a}(t)] \\ &\geq \frac{\eta_{\mathrm{att}}\gamma^6}{CM}\frac{s(t)^3}{(1+s(t))^3} \geq \frac{\eta_{\mathrm{att}}\gamma^6}{CM}. \end{aligned} \tag{197}$$

Equation equation 196 shows that the number of early Phase-II updates needed to reach $\overline{p} \geq 1/2$ is at most

$$\frac{CM}{\eta_{\mathrm{att}}\gamma^6}\max\{(M-1)e^{-(1-2c_0)\theta} - 1, 0\}. \tag{198}$$

Now put $z(t) = e^{2\overline{a}(t)}$ in the late regime. Since $e^z - 1 \geq z$,

$$\begin{aligned} z(t+1) - z(t) &\geq 2z(t)[\overline{a}(t+1) - \overline{a}(t)] \\ &\geq \frac{\eta_{\mathrm{att}}\gamma^6}{CM} z(t)\overline{p}(t)(1-\overline{p}(t))^2 \\ &= \frac{\eta_{\mathrm{att}}\gamma^6}{CM}\frac{1}{(1+e^{-\overline{a}(t)})^3} \geq \frac{\eta_{\mathrm{att}}\gamma^6}{CM}. \end{aligned} \tag{199}$$

After at most

$$\frac{CM}{\eta_{\mathrm{att}}\gamma^6\epsilon} \tag{200}$$

further iterations, $z(t) \geq C/\epsilon$, hence

$$(1-\overline{p}(t))^2 \leq \epsilon/C \tag{201}$$

. The sum of equation 198 and equation 200 is bounded by T_2 in equation 14.

For a prompt in the independently sampled terminal batch $\mathcal{B}_T$ having query class q, Lemma 6 and late synchronization give

$$\frac{1-p_n(t)}{p_n(t)} \leq e^{\rho + c_{\mathrm{sync}}} e^{-\overline{a}(t)}. \tag{202}$$

Thus $(1 - p_n(T))^2 \le C(1 - \overline{p}(T))^2$ uniformly over the terminal batch. Choosing the numerical target constant proves the first inequality in equation 16; exchanging channels proves the second. Loss result follows immediately. For the training loss

$$\|\nu_q - u_1\|_2^2 = \gamma^2((1 - p_n)^2 + \sum_{a \neq q} \pi_a^2) \le 2\gamma^2(1 - p_n)^2, \tag{203}$$

$$\|f_q - v_2\|_2^2 \le 2\gamma^2(1 - q_n)^2. \tag{204}$$

Since $\alpha = (1, 0)$ and $\beta = (0, 1)$,

$$\frac{1}{2}\|y_q - \widehat{y}_q\|_2^2 \le \gamma^2 \left[(1 - p_n)^2 + (1 - q_n)^2\right] \le 2\gamma^2\epsilon. \tag{205}$$

By the probability ledger, the handoff and all Phase-II events fail with probability at most $3\delta/4$, which is smaller than δ.

A.17.2 Compositional Generalization

For test prompt P_{test} let q be the index of its μ pattern. Define

$$\begin{aligned} a_{\text{test}} &= \frac{1 - p_{\text{test}}}{p_{\text{test}}} \\ &= \frac{\sum_{r \neq q} e^{x_\ell^\top W_1(t) x_q}}{\sum_{\ell = q} e^{x_\ell^\top W_1(t) x_q}} \\ &\overset{(1)}{\lesssim} (M - 1) e^{-\log \frac{(M-1)^2}{\epsilon}} \\ &< \epsilon \end{aligned} \tag{206}$$

Step (1) follows since in convergence proof we showed the logit difference is governed by the difference of μ patterns for head 1 and after convergence this difference becomes arbitrarily large. Finally, with value specialization, i.e. $\alpha = (1, 0)^\top, \mathbf{beta} = (0, 1)^\top$, we have

$$\begin{aligned} r_\nu &= \nu_q - \alpha_1 u_1 - \alpha_2 u_2 \\ &= \nu_q - \alpha_1 u_1 \\ &= \nu_q - u_1 \\ \|r_\nu\|^2 &= \|\nu_q - p_{\text{test}} \nu_q - \sum_{i \neq q} \pi_i \nu_i\|^2 \\ &= \gamma^2 \Big((1 - p_{\text{test}})^2 + \sum_{i \neq q} \pi_i^2 \Big) \\ &\le 2\gamma^2 (1 - p_{\text{test}})^2 \\ &< 2\gamma^2 \epsilon \end{aligned} \tag{207}$$

Proof for head 2 is analogous. Hence equation 15 is proven.

A.18 Proposition 1

Part (i) follows directly from the uniform terminal attention bound equation 201, since the preferred class mass $p_n(T)$ (resp. $q_n(T)$) is exactly the total attention assigned by head 1 (resp. head 2) to context tokens whose μ-pattern (resp. e-pattern) matches that of the query.

For part (ii), let $A_\mu(T) = P_\mu^\top W_1(T) P_\mu$. For every query class i, the clean log-odds identity gives

$$a_i(T) = (A_\mu(T))_{i,i} - \log \sum_{j \neq i} \exp((A_\mu(T))_{j,i}) \,. \tag{208}$$

The terminal attention guarantee together with the clean-to-prompt comparison implies

$$a_i(T) \geq \log \frac{1-\sqrt{\epsilon}}{\sqrt{\epsilon}} - O(1). \tag{209}$$

The terminal wrong-score tube implies that all wrong entries in column i differ by at most $O(1)$. Therefore, for every $j \neq i$,

$$(A_\mu(T))_{i,i} - (A_\mu(T))_{j,i} \geq \log \frac{M-1}{\sqrt{\epsilon}} - O(1). \tag{210}$$

The corresponding statement for $P_e^\top W_2(T) P_e$ follows by exchanging the two channels. The centered non-preferred block bounds follow directly from the terminal leakage bound.

Finally, part (iii) is exactly the value-specialization conclusion equation 132.

A.19 Proof of Out-of-domain generalization Theorem 2

We prove the μ'-channel bound. The e'-channel follows by exchanging the two channels.

Let

$$A_\mu := P_\mu^\top W_1(T) P_\mu. \tag{211}$$

For every in-domain query class q, define

$$b_q := \frac{1}{M-1} \sum_{s \neq q} (A_\mu)_{s,q}, \qquad \Delta_q := (A_\mu)_{q,q} - b_q. \tag{212}$$

The terminal wrong-score invariant implies

$$|(A_\mu)_{s,q} - b_q| \leq c_{\mathrm{wrong}}, \qquad s \neq q. \tag{213}$$

Set

$$\phi_q := \log \left[\frac{1}{M-1} \sum_{s \neq q} \exp\big((A_\mu)_{s,q} - b_q\big) \right]. \tag{214}$$

The arguments in equation 214 have average zero. Hence Jensen's inequality and equation 213 give

$$0 \leq \phi_q \leq c_{\mathrm{wrong}}. \tag{215}$$

If $a_q(T)$ is the terminal clean log odds, then

$$\Delta_q = a_q(T) + \log(M-1) + \phi_q. \tag{216}$$

The in-domain terminal evaluation batch is query-balanced. Hence each query class occurs in that batch. On $\mathcal{E}_{\mathrm{ret}}$, the corresponding preferred attention mass is at least $1-\sqrt{\epsilon}$, and the clean-to-prompt comparison gives

$$a_q(T) \geq \log \left(\frac{1-\sqrt{\epsilon}}{\sqrt{\epsilon}} \right) - \rho. \tag{217}$$

Combining equation 216, equation 215, and equation 217,

$$\Delta_* := \min_q \Delta_q \geq \log \left(\frac{(M-1)(1-\sqrt{\epsilon})}{\sqrt{\epsilon}} \right) - \rho. \tag{218}$$

Terminal synchronization also gives a fixed constant C_{tr}, independent of M and ϵ, such that

$$\max_{q,r} |\Delta_q - \Delta_r| \leq C_{\mathrm{tr}}. \tag{219}$$

Write exactly

$$A_\mu = \mathbf{1} b^\top + \Delta_* I_M + D_\mu + E_\mu, \tag{220}$$

where D_μ is diagonal and E_μ has zero diagonal. From equation 213 and equation 219,

$$\|D_\mu\|_{\mathrm{op}} \leq C_{\mathrm{tr}}, \qquad \|E_\mu\|_{\max} \leq c_{\mathrm{wrong}}. \tag{221}$$

Fix distinct OOD classes a, b. Since all columns of C_μ have the same sum, the gauge term in equation 220 cancels:

$$(c_a^\mu - c_b^\mu)^\top \mathbf{1} b^\top c_a^\mu = 0. \tag{222}$$

Therefore

$$\begin{aligned}(c_a^\mu - c_b^\mu)^\top A_\mu c_a^\mu = \Delta_* \left[1 - (c_b^\mu)^\top c_a^\mu\right] \\ + (c_a^\mu - c_b^\mu)^\top D_\mu c_a^\mu + (c_a^\mu - c_b^\mu)^\top E_\mu c_a^\mu.\end{aligned} \tag{223}$$

By definition,

$$1 - (c_b^\mu)^\top c_a^\mu \geq 1 - \rho_{\mathrm{ood}}. \tag{224}$$

The last two terms satisfy

$$\left|(c_a^\mu - c_b^\mu)^\top D_\mu c_a^\mu\right| \leq 2C_{\mathrm{tr}}, \tag{225}$$

and

$$\left|(c_a^\mu - c_b^\mu)^\top E_\mu c_a^\mu\right| \leq 2K_{\mathrm{ood}}^2 c_{\mathrm{wrong}}. \tag{226}$$

It remains to transfer the learned non-signal terms to the OOD patterns. The leakage proof works in the original coefficient coordinates of the fifteen centered non-signal blocks; before conversion to token scores, its stopped recursion and martingale bound give

$$\max |\text{terminal centered coefficient coordinate}| \leq C_{\mathrm{coef}} c_{\mathrm{leak}}, \tag{227}$$

where C_{coef} depends only on the fixed conversion constants of the convergence proof. Consider any one of those fifteen already-defined blocks. Substituting an OOD context pattern and an OOD query pattern into that block produces a bilinear sum of its coefficient coordinates. By the triangle inequality, its absolute contribution is at most $C_{\mathrm{coef}} c_{\mathrm{leak}}$ times the product of the corresponding context- and query-side coefficient ℓ_1-norms. For a main-family coefficient vector, centering changes the ℓ_1-norm by at most a factor two:

$$\left\| c - \frac{\mathbf{1}^\top c}{M} \mathbf{1} \right\|_1 \leq 2\|c\|_1. \tag{228}$$

Together with OOD matrices admissibility, the corresponding bounds for C_{z_μ}, C_{z_e}, and the perturbation-amplitude bound c, the preceding bilinear bound shows that each retained block contributes at most a fixed multiple of $K_{\mathrm{ood}}^2 c_{\mathrm{leak}}$. Since the number of retained blocks is exactly fifteen,

$$\left|r_{\ell,1}^{\mathrm{learn}}\right| \leq C_{\mathrm{mix}} K_{\mathrm{ood}}^2 c_{\mathrm{leak}}, \tag{229}$$

for a fixed constant C_{mix} independent of M and ϵ.

The only fixed non-clean initialization term of head 1 is the z_μ-z_μ block. The initialization formula, the perturbation amplitude bound, and the ℓ_2-bound on the columns of C_{z_μ} give

$$\left|r_{\ell,1}^{\mathrm{fix}}\right| \leq \theta c^2 K_{\mathrm{ood}}^2. \tag{230}$$

Thus every head-1 OOD nuisance score satisfies

$$|r'_{\ell,1}| \leq K_{\mathrm{nuis}}, \qquad K_{\mathrm{nuis}} := C_{\mathrm{mix}} K_{\mathrm{ood}}^2 c_{\mathrm{leak}} + \theta c^2 K_{\mathrm{ood}}^2. \tag{231}$$

Let $N_a^{\mu'}$ be the number of context tokens in OOD class a. For a query of class a, the nuisance part of the class-b log-partition therefore satisfies

$$\log N_b^{\mu'} - K_{\mathrm{nuis}} \leq \log \sum_{\ell:\, a_\ell = b} e^{r'_{\ell,1}} \leq \log N_b^{\mu'} + K_{\mathrm{nuis}}. \tag{232}$$

For a fixed $0 < \vartheta_{\mathrm{ctx}} < 1$ such that, simultaneously for every OOD prompt and every main class,

$$(1 - \vartheta_{\mathrm{ctx}}) \frac{L}{M} \leq N_a^{\mu'},\ N_i^{e'} \leq (1 + \vartheta_{\mathrm{ctx}}) \frac{L}{M}. \tag{233}$$

On the count event equation 233,

$$\left| \log \frac{N_b^{\mu'}}{N_a^{\mu'}} \right| \leq \log \frac{1 + \vartheta_{\mathrm{ctx}}}{1 - \vartheta_{\mathrm{ctx}}}. \tag{234}$$

Consequently, the nuisance log-partitions of any two OOD μ'-classes differ by at most

$$2K_{\mathrm{nuis}} + \log \frac{1+\vartheta_{\mathrm{ctx}}}{1-\vartheta_{\mathrm{ctx}}}, \tag{235}$$

which is a fixed constant independent of M and ϵ.

Combining equation 218, equation 223, equation 224, equation 225, equation 226, and equation 235, there is a fixed constant C_μ, independent of M and ϵ, such that every wrong OOD μ'-class has full class log-mass below the preferred class by at least

$$(1-\rho_{\mathrm{ood}}) \log \left(\frac{(M-1)(1-\sqrt{\epsilon})}{\sqrt{\epsilon}} \right) - C_\mu. \tag{236}$$

Therefore

$$\frac{1-p'_n(T)}{p'_n(T)} \le e^{C_\mu} (M-1)^{\rho_{\mathrm{ood}}} \left(\frac{\sqrt{\epsilon}}{1-\sqrt{\epsilon}} \right)^{1-\rho_{\mathrm{ood}}}. \tag{237}$$

The channel-exchanged argument gives a fixed constant C_e such that

$$\frac{1-q'_n(T)}{q'_n(T)} \le e^{C_e} (M-1)^{\rho_{\mathrm{ood}}} \left(\frac{\sqrt{\epsilon}}{1-\sqrt{\epsilon}} \right)^{1-\rho_{\mathrm{ood}}}. \tag{238}$$

On $\mathcal{E}_{\mathrm{ret}}$, the returned model has exact value specialization,

$$\alpha(T) = (1,0)^\top, \qquad \beta(T) = (0,1)^\top. \tag{239}$$

Hence

$$\widehat{y}'_q = \begin{bmatrix} \sum_b \pi'_b \nu'_b \\ \sum_j \rho'_j f'_j \end{bmatrix}. \tag{240}$$

Using OOD admissibility assumptions, and orthogonality of the original output patterns,

$$\left\| \nu'_a - \sum_b \pi'_b \nu'_b \right\|_2^2 \le 2\gamma^2 K_{\mathrm{ood}}^2 (1-p'_n)^2, \tag{241}$$

$$\left\| f'_i - \sum_j \rho'_j f'_j \right\|_2^2 \le 2\gamma^2 K_{\mathrm{ood}}^2 (1-q'_n)^2. \tag{242}$$

Using $x/(1+x) \le x$ in equation 237–equation 238, squaring, and absorbing all fixed factors into one constant C_{ood} proves equation 21.

A.20 Technical proof of Phase-I exchange stability

Fix two distinct μ-classes q, q' and exchange these two labels simultaneously in every μ-indexed row and column of the probed blocks of W_1 and W_2. Define

$$d_{q,q'}(t) := \max_{\text{probed coefficients}} \left| \text{coefficient at time } t - \text{its } q \leftrightarrow q' \text{ relabeling} \right|. \tag{243}$$

By the complementary initialization,

$$d_{q,q'}(0) = 0. \tag{244}$$

The balanced query law is invariant under the same relabeling. For a fixed task, exchanging two μ-labels also exchanges their corresponding output images under g_μ. Since the output patterns are orthogonal with equal norm and the value maps are scalar on each output channel, the squared loss and the attention-gradient formula are unchanged by this simultaneous exchange. Hence there is no population exchange forcing when $d_{q,q'}(t) = 0$.

Inside the Phase-I score tube, every main-class attention mass is at most C/M. For one prompt, write

$$g_\ell = \sigma_\ell \Gamma_\ell.$$

The softmax derivative gives

$$\frac{\partial g_\ell}{\partial s_r} = \sigma_\ell(\mathbf{1}\{\ell = r\} - \sigma_r)\Gamma_\ell + \sigma_\ell \frac{\partial \Gamma_\ell}{\partial s_r}. \tag{245}$$

Since $|\Gamma_\ell| \le 8$ and

$$\left\| \frac{\partial u_h}{\partial s_r} \right\|_2, \left\| \frac{\partial v_h}{\partial s_r} \right\|_2 \le 2\gamma\sigma_r,$$

the exact expression for Γ_ℓ gives

$$\left| \frac{\partial \Gamma_\ell}{\partial s_r} \right| \le C\sigma_r.$$

Therefore

$$\sum_r \left| \frac{\partial g_\ell}{\partial s_r} \right| \le C\sigma_\ell, \qquad \sum_\ell \left| \frac{\partial g_\ell}{\partial s_r} \right| \le C\sigma_r. \tag{246}$$

Summing equation 246 over one main class contributes a factor C/M, and exact query balance contributes a second factor $1/M$. Perturbation coordinates are no larger because

$$\mathbb{E}[\kappa^2 \mathbf{1}\{z = z_k\}] \le \frac{C}{M_z} \le \frac{C}{M}.$$

For a fixed exchange $q \leftrightarrow q'$ and a fixed probed coefficient, let $\Delta(t)$ denote the signed difference between that coefficient and its $q \leftrightarrow q'$ relabeling. Thus

$$|\Delta(t)| \le d_{q,q'}(t), \qquad d_{q,q'}(t) = \max_{\text{probed coefficients}} |\Delta(t)|. \tag{247}$$

The exchange symmetry proved above gives $\Delta(0) = 0$ and, when all exchanged coefficients agree, equality of the corresponding conditional expected updates. Applying the mean-value theorem between the current parameter state and its relabeling, together with equation 246, gives

$$\max_{\text{probed coefficients}} |\mathbb{E}_t[\Delta(t+1)]| \le \left(1 + \frac{C\eta_{\text{att}}\gamma^6}{M^2}\right) d_{q,q'}(t), \tag{248}$$

where $\mathbb{E}_t$ denotes expectation over the fresh batch at iteration t conditional on the preceding batches.

Consequently, a centered fluctuation generated at iteration s can change any fixed probed coefficient at a later iteration $t > s$ by at most the deterministic amplification factor

$$\left(1 + \frac{C\eta_{\text{att}}\gamma^6}{M^2}\right)^{t-1-s} \le \exp\left(\frac{C\eta_{\text{att}}\gamma^6 T_1}{M^2}\right). \tag{249}$$

From the exact attention update, the centered fresh-batch contribution to one fixed exchanged coefficient has Bernstein range

$$\frac{C\eta_{\text{att}}\gamma^6}{BM} \tag{250}$$

and conditional variance proxy

$$\frac{C\eta_{\text{att}}^2\gamma^{12}}{BM^3}. \tag{251}$$

For cross-channel query coordinates, the labels within each balanced block are sampled without replacement; the Bernstein inequality for uniformly permuted sums gives the same two bounds, up to numerical constants.

Fix now $t \le T_1$ and one probed coefficient. Expose the fresh batches from iterations $0, \dots, t-1$ successively. By equation 249, the centered contribution created at iteration s has conditional variance proxy at time t at most

$$\left(1 + \frac{C\eta_{\text{att}}\gamma^6}{M^2}\right)^{2(t-1-s)} \frac{C\eta_{\text{att}}^2\gamma^{12}}{BM^3}, \tag{252}$$

and Bernstein range at most

$$\left(1+\frac{C\eta_{\mathrm{att}}\gamma^6}{M^2}\right)^{t-1-s}\frac{C\eta_{\mathrm{att}}\gamma^6}{BM}. \tag{253}$$

Hence, uniformly for $t \le T_1$,

$$\sum_{s=0}^{t-1}\left(1+\frac{C\eta_{\mathrm{att}}\gamma^6}{M^2}\right)^{2(t-1-s)}\frac{C\eta_{\mathrm{att}}^2\gamma^{12}}{BM^3} \le C\exp\left(\frac{2C\eta_{\mathrm{att}}\gamma^6T_1}{M^2}\right)\frac{\eta_{\mathrm{att}}^2\gamma^{12}T_1}{BM^3}, \tag{254}$$

$$\max_{0\le s<t}\left(1+\frac{C\eta_{\mathrm{att}}\gamma^6}{M^2}\right)^{t-1-s}\frac{C\eta_{\mathrm{att}}\gamma^6}{BM} \le C\exp\left(\frac{C\eta_{\mathrm{att}}\gamma^6T_1}{M^2}\right)\frac{\eta_{\mathrm{att}}\gamma^6}{BM}. \tag{255}$$

The Bernstein–Freedman maximal inequality applied to this batch-exposure martingale therefore gives, for every $u > 0$,

$$\mathbb{P}(|\Delta(t)| \ge u) \le 2\exp\left(-\frac{u^2}{2\left[Ce^{\,2C\eta_{\mathrm{att}}\gamma^6T_1/M^2}\dfrac{\eta_{\mathrm{att}}^2\gamma^{12}T_1}{BM^3} + Ce^{\,C\eta_{\mathrm{att}}\gamma^6T_1/M^2}\dfrac{\eta_{\mathrm{att}}\gamma^6}{BM}\,u\right]}\right). \tag{256}$$

There are at most $C(M+M_z)^2$ probed coefficients in both heads, fewer than M^2 exchanges $q \leftrightarrow q'$, and at most T_1+1 relevant times. Applying a union bound to equation 256 and solving the Bernstein bound with total failure probability δ_{exch} yields

$$\max_{q\ne q'}\max_{t\le T_1} d_{q,q'}(t) \le C\exp\left(\frac{C\eta_{\mathrm{att}}\gamma^6T_1}{M^2}\right)\left[\sqrt{\frac{\eta_{\mathrm{att}}^2\gamma^{12}T_1}{BM^3}\log\frac{C(M+M_z)^2M^2(T_1+1)}{\delta_{\mathrm{exch}}}} + \frac{\eta_{\mathrm{att}}\gamma^6}{BM}\log\frac{C(M+M_z)^2M^2(T_1+1)}{\delta_{\mathrm{exch}}}\right]. \tag{257}$$

By equation 75,

$$\max_{q\ne q'}\max_{t\le T_1} d_{q,q'}(t) \le \frac{c_{\mathrm{sync}}}{C_{\mathrm{hand}}M}. \tag{258}$$

A.21 Technical proof of leakage stability

This appendix proves Lemma 9. We prove the statement for head 1; head 2 follows by exchanging the μ- and e-channels. Throughout the proof we work on the stopped interval

$$T_1 \le t < \tau,$$

where τ is the Phase-II stopping time defined in Section 4.2. Thus value specialization, preferred-class maximality, the wrong-score bound, leakage control, synchronization, and the KKT conditions may all be used below.

Exact representative updates. For the query-side μ–e coordinate (a,i), the exact update is

$$\frac{\eta_{\mathrm{att}}\gamma^6}{B}\sum_{n=1}^{B}h_a^{(n)}(t)\left(\mathbf{1}\{e_q^{(n)}=e_i\}-\frac{1}{M}\right). \tag{259}$$

Replacing the parenthesis by $\kappa_{\mu,q}^{(n)}\mathbf{1}\{z_{\mu,q}^{(n)}=z_{\mu,k}\}$ or $\kappa_{e,q}^{(n)}\mathbf{1}\{z_{e,q}^{(n)}=z_{e,k}\}$ gives the other two query-side blocks.

For the context-side e–μ coordinate (i,q), the exact update is

$$\frac{\eta_{\mathrm{att}}\gamma^6}{B}\sum_{n:j_n=q}\sum_{\ell=1}^{L}\sigma_{\ell,1}^{(n)}(t)\Gamma_{\ell,1}^{(n)}(t)\left(\mathbf{1}\{e_\ell^{(n)}=e_i\}-\frac{1}{M}\right), \tag{260}$$

where $|\Gamma^{(n)}_{\ell,1}| \le 8$. Replacing the centered context selector gives the other two context-side blocks. Multiplying the summand in equation 260 by one centered query selector gives the nine two-sided blocks. These formulas cover all fifteen blocks.

When a nuisance main selector is written as centered part plus its mean, the mean-context terms are constant across all context tokens and are therefore exact softmax gauge terms. They have zero attention gradient. We retain only the nongauge centered terms described above.

A conditional finite-context cancellation lemma. Fix one prompt at iteration t, condition on its query including all query nuisance variables, and condition on the vector of main-class assignments of its L context positions. Let N_a be the number of context positions whose main class is a. By the same multiplicative Chernoff argument used in Lemma 6, the event in equation 171 may be chosen so that, simultaneously for all prompts used through time T,

$$\frac{L}{2M} \le N_a \le \frac{2L}{M}, \qquad a = 1, \dots, M. \tag{261}$$

After this conditioning, nuisance variables belonging to different main classes are independent.

Finite-context cancellation for context-side leakage. Fix one prompt and condition on its query and on its main-class assignments. Let N_a be the number of context tokens in main class a. Consider either centered nuisance selector

$$\chi_\ell = \mathbf{1}\{e_\ell = e_i\} - \frac{1}{M} \qquad \text{or} \qquad \chi_\ell = \kappa_\ell \mathbf{1}\{z_\ell = z_k\}.$$

In both cases,

$$\mathbb{E}_t|\chi_\ell| \le \frac{C}{M}, \qquad \mathbb{E}_t\chi_\ell^2 \le \frac{C}{M}, \tag{262}$$

where the perturbation case uses $M_z \ge Cc^2M$.

Inside each main class, the clean and query-side factors are constant and cancel from the normalized class attention. The remaining nuisance softmax weights are bounded above and below by fixed constants on the stopped interval. Hence the softmax-weighted average of χ_ℓ over the N_a tokens in class a satisfies

$$\mathbb{E}_t\left[\left|\frac{\sum_{\ell:\mu_\ell=\mu_a} e^{z_{\ell,\mathrm{nuis}}}\chi_\ell}{\sum_{\ell:\mu_\ell=\mu_a} e^{z_{\ell,\mathrm{nuis}}}} - \frac{\mathbb{E}_t[e^{z_{\ell,\mathrm{nuis}}}\chi_\ell]}{\mathbb{E}_t[e^{z_{\ell,\mathrm{nuis}}}]}\right|^2\right] \le \frac{C}{MN_a} \le \frac{C}{L}. \tag{263}$$

Indeed, the denominator is bounded below by a fixed multiple of N_a, while the numerator is a sum of N_a centered bounded variables whose total conditional variance is at most CN_a/M.

We next record the only derivative estimates needed below. Recall

$$h_a = \pi_a\left(\mathbf{1}\{a = q\} - p - \pi_a + \sum_b \pi_b^2\right), \qquad p = \pi_q. \tag{264}$$

The wrong-score tube gives

$$\pi_s \le \frac{C(1-p)}{M}, \qquad s \ne q. \tag{265}$$

Using

$$\frac{\partial \pi_a}{\partial z_b} = \pi_a(\mathbf{1}\{a = b\} - \pi_b), \qquad \sum_b \left|\frac{\partial \pi_a}{\partial z_b}\right| \le 2\pi_a,$$

and differentiating equation 264 once and twice gives

$$\sum_b \left(\left|\frac{\partial h_q}{\partial z_b}\right| + \left|\frac{\partial^2 h_q}{\partial z_b^2}\right|\right) + \sum_{b,c}\left|\frac{\partial^2 h_q}{\partial z_b \partial z_c}\right| \le Cp(1-p)^2, \tag{266}$$

and, uniformly for $s \ne q$,

$$\sum_b \left(\left|\frac{\partial h_s}{\partial z_b}\right| + \left|\frac{\partial^2 h_s}{\partial z_b^2}\right|\right) + \sum_{b,c}\left|\frac{\partial^2 h_s}{\partial z_b \partial z_c}\right| \le \frac{C}{M}p(1-p)^2. \tag{267}$$

The same calculation also gives

$$\sum_a |h_a| + \sum_{a,b} \left| \frac{\partial h_a}{\partial z_b} \right| + \sum_{a,b,c} \left| \frac{\partial^2 h_a}{\partial z_b \partial z_c} \right| \le Cp(1-p)^2. \tag{268}$$

The factor $1/M$ in equation 267 follows directly from equation 265; thus summing over the $M-1$ wrong classes does not produce an additional factor M.

It remains to account for the fact that h_a and the within-class weighted nuisance average are computed from the same finite context. Expose the nuisance samples in class a one at a time and compare each sample with an independent copy having the same conditional mean. The first-order term is centered and therefore has zero conditional expectation. The Taylor remainder is bounded by the first- and second-derivative estimates equation 266–equation 268. Using equation **??** and equation 262, each preferred-class contribution is at most

$$\frac{C}{MN_q}\, p(1-p)^2,$$

and each wrong-class contribution is at most

$$\frac{C}{M^2 N_s}\, p(1-p)^2.$$

Summing over the main classes therefore gives

$$|\mathbb{E}_t[\text{one-prompt centered context-side contribution}]| \le \frac{C}{L} p(1-p)^2. \tag{269}$$

Similarly, using equation 263 and equation 268,

$$\mathbb{E}_t\Big[|\text{one-prompt centered context-side contribution}|^2\Big] \le \frac{C}{L} p^2(1-p)^4. \tag{270}$$

Multiplying by an additional centered query selector changes only the fixed numerical constant, so equation 269–equation 270 also apply to every retained two-sided leakage contribution.

Finite-context correction for the query-side blocks. For head 1, define the three query-side blocks by

$$Q_e(t) := P_\mu^\top W_1(t) P_e \left(I - \frac{1}{M} \mathbf{1}\mathbf{1}^\top \right), \tag{271}$$

and

$$Q_{z_\mu}(t) := P_\mu^\top W_1(t) P_{z_\mu}, \qquad Q_{z_e}(t) := P_\mu^\top W_1(t) P_{z_e}. \tag{272}$$

These are exactly the μ–e, μ–z_μ, and μ–z_e leakage blocks.

For a prompt whose μ-query class is q, let $h_{a,\infty}^{(n)}(t)$ denote the class-gradient coefficient obtained from the corresponding infinite-context class masses. Finite context changes these coefficients through two effects: fluctuations of the number of context examples in each main class and fluctuations of the nuisance variables inside each class.

Both effects can be treated by replacing the context samples one at a time. The first-order term has conditional mean zero because the class and nuisance fluctuations are centered. The second-order Taylor remainder is controlled by equation 266–equation 268. Using $N_a \asymp L/M$ therefore gives

$$\left| \mathbb{E}_t h_q^{(n)}(t) - h_{q,\infty}^{(n)}(t) \right| \le \frac{CM}{L}\, p_q(t)(1-p_q(t))^2, \tag{273}$$

and, uniformly for $a \neq q$,

$$\left| \mathbb{E}_t h_a^{(n)}(t) - h_{a,\infty}^{(n)}(t) \right| \le \frac{C}{L}\, p_q(t)(1-p_q(t))^2. \tag{274}$$

The extra factor $1/M$ in the wrong-class bound comes from equation 267. Hence summing over all wrong classes does not introduce an additional factor M.

We now apply these bounds to the three query-side updates. Exactly B/M prompts in every batch have each fixed μ-query class. Moreover, for every fixed query coordinate,

$$\mathbb{E}\left|\mathbf{1}\{e_q = e_i\} - \frac{1}{M}\right| \le \frac{C}{M}, \qquad \mathbb{E}\,|\kappa\,\mathbf{1}\{z = z_k\}| \le \frac{C}{M_z} \le \frac{C}{M}. \tag{275}$$

For a fixed row a, the stratum $q = a$ therefore contributes at most

$$\frac{1}{M}\,\frac{C}{M}\,\frac{CM}{L}\,p_a(t)(1-p_a(t))^2,$$

while each of the $M-1$ strata $q \neq a$ contributes at most

$$\frac{1}{M}\,\frac{C}{M}\,\frac{C}{L}\,p_q(t)(1-p_q(t))^2.$$

Summing the wrong-query terms gives, entrywise,

$$|\text{finite-context correction to the one-step update of } Q_e(t)| \le \frac{C\eta_{\text{att}}\gamma^6}{ML}\max_q p_q(t)(1-p_q(t))^2. \tag{276}$$

The identical calculation gives

$$\left|\text{finite-context correction to the one-step update of } Q_{z_\mu}(t)\right| \le \frac{C\eta_{\text{att}}\gamma^6}{ML}\max_q p_q(t)(1-p_q(t))^2, \tag{277}$$

and

$$|\text{finite-context correction to the one-step update of } Q_{z_e}(t)| \le \frac{C\eta_{\text{att}}\gamma^6}{ML}\max_q p_q(t)(1-p_q(t))^2. \tag{278}$$

Accumulated finite-context correction. The preceding bounds apply to every retained leakage coefficient. For the three query-side blocks, equation 276–equation 278 give, entrywise,

$$|\text{finite-context part of one update}| \le \frac{C\eta_{\text{att}}\gamma^6}{ML}\max_q p_q(t)(1-p_q(t))^2.$$

For the context-side and two-sided blocks, equation 269 gives the same order after the exact B/M query-stratum average is included.

By Lemma 7, the clean progress of every fixed query class satisfies

$$a_q(t+1) - a_q(t) \ge \frac{\eta_{\text{att}}\gamma^6}{CM}p_q(t)(1-p_q(t))^2$$

throughout the stopped interval. Synchronization makes $p_q(t)(1-p_q(t))^2$ comparable across query classes. Hence, for any fixed query class,

$$\sum_{t=T_1}^{T-1}\frac{\eta_{\text{att}}\gamma^6}{M}\max_r p_r(t)(1-p_r(t))^2 \le C\big[a_q(T) - a_q(T_1)\big]. \tag{279}$$

At the handoff, the clean score tube gives

$$a_q(T_1) = \theta - \log(M-1) + O(1).$$

Thus the increase before $p_q(t)$ reaches $1/2$ is at most $C\log M$.

After $p_q(t) \ge 1/2$, the clean-progress upper bound gives

$$a_q(t+1) - a_q(t) \le \frac{C\eta_{\text{att}}\gamma^6}{M}p_q(t)(1-p_q(t))^2 \le \frac{C\eta_{\text{att}}\gamma^6}{M}e^{-2a_q(t)}.$$

The attention step size is chosen so that $a_q(t+1) - a_q(t) \le 1$. Therefore

$$e^{2a_q(t+1)} - e^{2a_q(t)} \le Ce^{2a_q(t)}\big[a_q(t+1) - a_q(t)\big] \le \frac{C\eta_{\text{att}}\gamma^6}{M}.$$

Summing from the first time $p_q(t) \ge 1/2$ to T gives

$$a_q(T) - a_q(T_1) \le C\log(CMT).$$

Consequently,

$$\sum_{t=T_1}^{T-1}|\text{finite-context part of any retained one-step update}| \le \frac{C\log(CMT)}{L}. \tag{280}$$

Infinite-context evolution of the leakage blocks. We first show that the context-side and two-sided blocks have exactly zero infinite-context drift.

Fix the μ-query class q, the query form index i, and the query perturbation indices. After the clean and query-side terms are factored out, the remaining context nuisance variables have the same distribution in every main class. Therefore the normalized infinite-context class mass of class a is proportional to

$$\exp\big(A_{a,q}(t) + (Q_e(t))_{a,i} + \kappa_{\mu,q}(Q_{z_\mu}(t))_{a,k_\mu} + \kappa_{e,q}(Q_{z_e}(t))_{a,k_e}\big)\,.$$

Every context-side and two-sided coefficient contributes the same multiplicative factor to all main classes, and this factor cancels from the softmax normalization. Hence

$$\mathbb{E}_t[\text{one-step update of any context-side or two-sided coefficient}]_{\text{infinite context}} = 0. \tag{281}$$

It remains to study $Q_e(t)$, $Q_{z_\mu}(t)$, and $Q_{z_e}(t)$.

At the exactly synchronized clean profile with preferred mass $p(t)$ and zero query-side leakage, direct differentiation of the class-gradient coefficients gives

$$\frac{1}{M}\sum_{q=1}^{M}\left(-\frac{\partial h_a}{\partial z_b}\right) = \frac{(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{(M-1)^3}\left(I - \frac{1}{M}\mathbf{1}\mathbf{1}^\top\right)_{ab}. \tag{282}$$

For $p(t) \geq 1/2$,

$$\frac{(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{(M-1)^3} \geq \frac{p(t)(1-p(t))^2}{CM}. \tag{283}$$

For the centered e query coordinate,

$$\mathbb{E}\left[\left(\mathbf{1}\{e_q = e_i\} - \frac{1}{M}\right)\left(\mathbf{1}\{e_q = e_j\} - \frac{1}{M}\right)\right] = \frac{1}{M}\left(\mathbf{1}\{i=j\} - \frac{1}{M}\right). \tag{284}$$

For the perturbation query coordinates,

$$\mathbb{E}\left[\kappa_{\mu,q}^2\mathbf{1}\{z_{\mu,q} = z_{\mu,k}\}\mathbf{1}\{z_{\mu,q} = z_{\mu,k'}\}\right] = \frac{c^2}{3M_z}\mathbf{1}\{k=k'\}, \tag{285}$$

and the same identity holds for the z_e query coordinate. Mixed second moments between the e, z_μ, and z_e query coordinates are zero by independence and centering.

Every exact query-side update is an outer product of $(h_a^{(n)}(t))_{a=1}^M$ with the corresponding centered query coordinate. Since

$$\sum_a h_a^{(n)}(t) = 0,$$

the exact updates preserve

$$\mathbf{1}^\top Q_e(t) = \mathbf{1}^\top Q_{z_\mu}(t) = \mathbf{1}^\top Q_{z_e}(t) = 0. \tag{286}$$

Moreover, because the e query coordinate is centered,

$$Q_e(t)\mathbf{1} = 0. \tag{287}$$

Combining equation 282, equation 284, equation 285, equation 286, and equation 287, the infinite-context conditional expected updates at the exactly synchronized clean profile and zero query-side leakage are

$$\begin{aligned}
\mathbb{E}_t[Q_e(t+1) - Q_e(t)] &= -\eta_{\text{att}}\gamma^6\frac{(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{M(M-1)^3}\,Q_e(t),\\
\mathbb{E}_t[Q_{z_\mu}(t+1) - Q_{z_\mu}(t)] &= -\eta_{\text{att}}\gamma^6\frac{c^2(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{3M_z(M-1)^3}\,Q_{z_\mu}(t),\\
\mathbb{E}_t[Q_{z_e}(t+1) - Q_{z_e}(t)] &= -\eta_{\text{att}}\gamma^6\frac{c^2(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{3M_z(M-1)^3}\,Q_{z_e}(t).
\end{aligned} \tag{288}$$

We next compare these ideal updates with the actual infinite-context updates on the stopped interval. Since

$$-\frac{\partial h_a}{\partial z_b}$$

is the Hessian of the class-logit loss, its derivative with respect to a class logit is

$$-\frac{\partial^2 h_a}{\partial z_b \partial z_c}.$$

Using equation 266–equation 268 and exact query balance gives, for every fixed row,

$$\frac{1}{M}\sum_{q=1}^{M}\sum_{b,c}\left|\frac{\partial^2 h_a}{\partial z_b \partial z_c}\right| \le \frac{C}{M}p(t)(1-p(t))^2, \tag{289}$$

and the same estimate holds for every fixed column because the class-logit Hessian is symmetric.

The wrong-score tube and synchronization imply that the actual clean class logits differ, after removal of the common softmax gauge, from the exactly synchronized equal-wrong profile by at most

$$C(c_{\mathrm{wrong}} + c_{\mathrm{sync}})$$

in maximum norm. The stopped leakage condition adds at most c_{leak} to any query-side class logit. Integrating equation 289 along these two line segments therefore shows that the row and column sums of the actual averaged class-logit Hessian differ from those in equation 282 by at most

$$\frac{C}{M}(c_{\mathrm{wrong}} + c_{\mathrm{sync}} + c_{\mathrm{leak}})p(t)(1-p(t))^2. \tag{290}$$

The query coordinates also satisfy

$$\mathbb{E}\left|\mathbf{1}\{e_q = e_i\} - \frac{1}{M}\right| \le \frac{2}{M}, \qquad \mathbb{E}\sum_{i=1}^{M}\left|\mathbf{1}\{e_q = e_i\} - \frac{1}{M}\right| \le 2, \tag{291}$$

and

$$\mathbb{E}\,|\kappa\,\mathbf{1}\{z = z_k\}| \le \frac{C}{M_z}, \qquad \mathbb{E}\sum_{k=1}^{M_z}|\kappa\,\mathbf{1}\{z = z_k\}| \le C. \tag{292}$$

Hence equation 290 also controls every cross-block coupling between Q_e, Q_{z_μ}, and Q_{z_e}.

Choose the fixed radii so that

$$C(c_{\mathrm{wrong}} + c_{\mathrm{sync}} + c_{\mathrm{leak}}) \le \frac{1}{4}. \tag{293}$$

Suppose first that every $p_q(t) \ge 1/2$. For a row belonging to Q_e, the absolute row sum of its infinite-context one-step update is at most

$$\begin{aligned} &1 - \eta_{\mathrm{att}}\gamma^6\frac{(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{M(M-1)^3} \\ &\quad + \frac{1}{4}\eta_{\mathrm{att}}\gamma^6\frac{(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{M(M-1)^3} \le 1. \end{aligned} \tag{294}$$

The same estimate holds for its column sums.

For a row belonging to either Q_{z_μ} or Q_{z_e}, the corresponding bound is

$$\begin{aligned} &1 - \eta_{\mathrm{att}}\gamma^6\frac{c^2(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{3M_z(M-1)^3} \\ &\quad + \frac{1}{4}\eta_{\mathrm{att}}\gamma^6\frac{c^2(1-p(t))^2\left(3M^2p(t)^2 - 2Mp(t) + M - 2\right)}{3M_z(M-1)^3} \le 1, \end{aligned} \tag{295}$$

and again the same estimate holds for columns.

Thus, if only the infinite-context conditional expected update is applied and every $p_q(t) \geq 1/2$,

$$\begin{aligned}\max\Big\{\|Q_e(t+1)\|_{\max}, \|Q_{z_\mu}(t+1)\|_{\max}, \|Q_{z_e}(t+1)\|_{\max}\Big\} \\ \leq \max\Big\{\|Q_e(t)\|_{\max}, \|Q_{z_\mu}(t)\|_{\max}, \|Q_{z_e}(t)\|_{\max}\Big\}.\end{aligned} \tag{296}$$

While some $p_q(t) < 1/2$, we do not use the favorable sign of equation 282. The same derivative bounds, exact query balance, equation 291, and equation 292 give

$$\begin{aligned}\max\Big\{\|Q_e(t+1)\|_{\max}, \|Q_{z_\mu}(t+1)\|_{\max}, \|Q_{z_e}(t+1)\|_{\max}\Big\} \\ \leq \left(1 + \frac{C\eta_{\mathrm{att}}\gamma^6}{M^2} \max_q p_q(t)(1-p_q(t))^2\right) \\ \times \max\Big\{\|Q_e(t)\|_{\max}, \|Q_{z_\mu}(t)\|_{\max}, \|Q_{z_e}(t)\|_{\max}\Big\},\end{aligned} \tag{297}$$

again for the infinite-context conditional expected update alone.

The derivative estimates used above hold uniformly along every line segment inside the stopped leakage tube. Hence equation 296–equation 297 also control the propagation of any perturbation introduced at an intermediate iteration.

By Lemma 7, synchronization, and equation 279,

$$\sum_{\{t:\min_q p_q(t)<1/2\}} \frac{\eta_{\mathrm{att}}\gamma^6}{M^2} \max_q p_q(t)(1-p_q(t))^2 \leq \frac{C\log M}{M}. \tag{298}$$

Therefore any leakage perturbation introduced at any time is amplified by all subsequent infinite-context population updates by at most

$$\prod_{\{t:\min_q p_q(t)<1/2\}} \left(1 + \frac{C\eta_{\mathrm{att}}\gamma^6}{M^2} \max_q p_q(t)(1-p_q(t))^2\right) \leq \exp\left(\frac{C\log M}{M}\right) \leq C_{\mathrm{prod}}. \tag{299}$$

The context-side and two-sided coefficients have no infinite-context drift by equation 281, so the same bound trivially applies to them.

Combining equation 280 with equation 299, and enlarging only the numerical constant in equation 171, gives

$$C_{\mathrm{prod}} \sum_{t=T_1}^{T-1} |\text{finite-context part of any retained one-step update}| \leq \frac{c_{\mathrm{leak}}}{8}. \tag{300}$$

Fresh-batch fluctuations. For an entry of $Q_e(t)$, exact query balance, equation 266–equation 267, and the second moment in equation 284 give

$$\begin{aligned}\mathbb{E}_t\left[\left|(Q_e(t+1)-Q_e(t))_{a,i} - \mathbb{E}_t[(Q_e(t+1)-Q_e(t))_{a,i}]\right|^2\right] \\ \leq \frac{C\eta_{\mathrm{att}}^2\gamma^{12}}{BM^2}\left[\max_q p_q(t)(1-p_q(t))^2\right]^2.\end{aligned} \tag{301}$$

Using equation 285 gives the same upper bound for every entry of $Q_{z_\mu}(t)$ and $Q_{z_e}(t)$, because $M_z \geq Cc^2M$.

For every retained context-side or two-sided coefficient, equation 270 and exact query balance give

$$\begin{aligned}\mathbb{E}_t\left[\left|\text{one-step coefficient update} - \mathbb{E}_t[\text{one-step coefficient update}]\right|^2\right] \\ \leq \frac{C\eta_{\mathrm{att}}^2\gamma^{12}}{BML}\left[\max_q p_q(t)(1-p_q(t))^2\right]^2.\end{aligned} \tag{302}$$

Lemma 8 therefore gives, uniformly over every retained coefficient,

$$\sum_{t=T_1}^{T-1} \mathbb{E}_t\left[\left|\text{one-step coefficient update} - \mathbb{E}_t[\text{one-step coefficient update}]\right|^2\right] \le C\eta_{\mathrm{att}}\gamma^6\left(\frac{1}{BM}+\frac{1}{BL}\right). \tag{303}$$

Moreover, one prompt changes any retained coefficient update by at most

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{B}. \tag{304}$$

We now control these centered fresh-batch fluctuations uniformly over time. The mean-value argument used in equation 296–equation 297 shows that a fluctuation inserted at any iteration is subsequently amplified by at most C_{prod}. The same statement is trivial for context-side and two-sided coefficients because their infinite-context drift is zero.

Apply the conditional Bernstein inequality to the centered fresh-batch increments and use the row- and column-sum bounds proved above. The standard Bernstein–Ville argument obtained by summing the $\cosh$ potentials over all retained coefficients gives

$$\begin{aligned}\mathbb{P}\left(\sup_{T_1\le t\le T}\max_{\text{retained coefficients}}\left|\text{accumulated propagated centered fresh-batch contribution}\right| > \frac{c_{\mathrm{leak}}}{8C}\right)\\ \le C(M+M_z)^2\exp\left[-\frac{c_{\mathrm{leak}}^2}{CC_{\mathrm{prod}}^2\left(\eta_{\mathrm{att}}\gamma^6\left(\frac{1}{BM}+\frac{1}{BL}\right)+\frac{\eta_{\mathrm{att}}\gamma^6 c_{\mathrm{leak}}}{B}\right)}\right].\end{aligned} \tag{305}$$

There are only $C(M+M_z)^2$ retained coefficients. Hence, after increasing only the numerical constant in equation 171, the right-hand side of equation 305 is at most $\delta/8$.

By the definition of the centered leakage score, each score is a sum of only a fixed number of retained block contributions, and every categorical or perturbation multiplier is bounded by a fixed constant depending only on c. Thus the event in equation 305 implies that the total propagated fresh-batch contribution to every centered leakage score is at most $c_{\mathrm{leak}}/8$.

Finally, the Phase-I handoff gives

$$\max\left|\text{centered leakage score at } T_1\right| \le \frac{c_{\mathrm{leak}}}{4C_{\mathrm{prod}}}.$$

By equation 299, its contribution at every later time is therefore at most $c_{\mathrm{leak}}/4$. The propagated finite-context contribution is at most $c_{\mathrm{leak}}/8$ by equation 300, and the propagated centered fresh-batch contribution is at most $c_{\mathrm{leak}}/8$ with probability at least $1-\delta/8$. Consequently,

$$\max\left|\text{centered leakage score}\right| \le \frac{c_{\mathrm{leak}}}{4}+\frac{c_{\mathrm{leak}}}{8}+\frac{c_{\mathrm{leak}}}{8} = \frac{c_{\mathrm{leak}}}{2} < c_{\mathrm{leak}}$$

throughout $T_1 \le t \le T$.

Therefore the leakage boundary cannot be the first boundary reached by the Phase-II stopped process. The head 2 argument is identical after exchanging the μ- and e-channels. This proves Lemma 9.

A.22 Technical proof of synchronization persistence

This appendix proves the synchronization lemma. We prove the result for head 1; head 2 follows by exchanging the μ- and e-channels.

Proof. We treat separately the regime before and after all clean preferred masses reach $1/2$.

Early regime. Fix two distinct μ-query classes q and q'. Exchange the labels q and q' throughout the μ-channel. The balanced query design is invariant under this relabeling, and the fresh batches can therefore be coupled by the same exchange at every iteration.

Throughout this part of the proof we work on the stopped interval on which

$$\left|\frac{1-p_q(t)}{p_q(t)}-\frac{1-p_{q'}(t)}{p_{q'}(t)}\right|\le\frac{c_{\mathrm{sync}}}{2},$$

and before all clean preferred masses have reached $1/2$.

For a wrong class $s\neq q$, the clean coefficient that must be controlled under the exchange is

$$A_{s,q}(t)-\frac{1}{M-1}\sum_{r\neq q}A_{r,q}(t). \tag{306}$$

This removes the common preferred-versus-average-wrong direction and keeps only the shape of the wrong clean scores.

Using $\sum_r h_r^{(n)}(t)=0$, its exact update is

$$\begin{aligned}&\left[A_{s,q}(t+1)-\frac{1}{M-1}\sum_{r\neq q}A_{r,q}(t+1)\right]-\left[A_{s,q}(t)-\frac{1}{M-1}\sum_{r\neq q}A_{r,q}(t)\right]\\&\quad=\frac{\eta_{\mathrm{att}}\gamma^6}{B}\sum_{n:j_n=q}\left[h_s^{(n)}(t)+\frac{h_q^{(n)}(t)}{M-1}\right].\end{aligned} \tag{307}$$

For the query-side leakage, we compare directly the corresponding entries of the three blocks

$$Q_e(t),\qquad Q_{z_\mu}(t),\qquad Q_{z_e}(t)$$

defined in Appendix A.21. The context-side and two-sided leakage blocks do not have an infinite-context class-dependent effect, as proved in Appendix A.21; they enter the present argument only through their finite-context corrections.

We first bound how differences in equation 306 and in the corresponding entries of Q_e, Q_{z_μ}, Q_{z_e} can grow under the infinite-context conditional updates.

At an equal-wrong clean profile with preferred mass p, direct differentiation of

$$h_s+\frac{h_q}{M-1}$$

in a centered wrong-score direction gives the factor

$$2\frac{(1-p)^2}{(M-1)^2}+\frac{(1-p)^2}{M-1}\left(p-\frac{1-p}{M-1}\right). \tag{308}$$

Throughout Phase II, $p\ge 1/(CM)$, and therefore

$$\left|2\frac{(1-p)^2}{(M-1)^2}+\frac{(1-p)^2}{M-1}\left(p-\frac{1-p}{M-1}\right)\right|\le\frac{C}{M}p(1-p)^2. \tag{309}$$

The wrong-score and leakage tubes change this estimate only by a fixed factor. Since exactly B/M prompts have query class q, exact query balance gives one further factor $1/M$. Hence the contribution of the clean wrong-score shape to one-step amplification is at most

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{M^2}\max_r p_r(t)(1-p_r(t))^2. \tag{310}$$

For $Q_e(t)$, the centered query coordinate satisfies

$$\mathbb{E}\left[\left(\mathbf{1}\{e_q=e_i\}-\frac{1}{M}\right)\left(\mathbf{1}\{e_q=e_j\}-\frac{1}{M}\right)\right]=\frac{1}{M}\left(\mathbf{1}\{i=j\}-\frac{1}{M}\right). \tag{311}$$

For $Q_{z_\mu}(t)$ and $Q_{z_e}(t)$,

$$\mathbb{E}\left[\kappa^2\mathbf{1}\{z=z_k\}\mathbf{1}\{z=z_{k'}\}\right]=\frac{c^2}{3M_z}\mathbf{1}\{k=k'\}\le\frac{C}{M}\mathbf{1}\{k=k'\}, \tag{312}$$

where $M_z \geq Cc^2 M$ was used.

Combining these second moments with the derivative bounds for the class-gradient coefficients and exact query balance gives the same one-step bound

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{M^2} \max_r p_r(t)(1-p_r(t))^2 \tag{313}$$

for the corresponding entries of Q_e, Q_{z_μ}, Q_{z_e}.

The clean wrong-score coefficients and the three query-side blocks can also affect each other. At zero query-side leakage these cross effects vanish exactly because

$$\mathbb{E}\left[\mathbf{1}\{e_q = e_i\} - \frac{1}{M}\right] = 0, \qquad \mathbb{E}\left[\kappa \mathbf{1}\{z = z_k\}\right] = 0.$$

At the actual query-side leakage, the mean-value theorem and the second-derivative bounds for h_a show that each cross effect is at most

$$\frac{Cc_{\mathrm{leak}}\eta_{\mathrm{att}}\gamma^6}{M^2} \max_r p_r(t)(1-p_r(t))^2. \tag{314}$$

Therefore, if two coupled states differ only in their centered clean wrong-score coefficients and in corresponding entries of Q_e, Q_{z_μ}, Q_{z_e}, their coefficient differences can increase in one infinite-context step by at most the factor

$$1 + \frac{C\eta_{\mathrm{att}}\gamma^6}{M^2} \max_r p_r(t)(1-p_r(t))^2. \tag{315}$$

Appendix A.21 already proves

$$\sum_{\{t:\min_r p_r(t)<1/2\}} \frac{\eta_{\mathrm{att}}\gamma^6}{M^2} \max_r p_r(t)(1-p_r(t))^2 \leq \frac{C\log M}{M}. \tag{316}$$

Hence every exchanged coefficient difference is amplified by the subsequent infinite-context dynamics by at most C_{prod}.

There is one additional population effect: $a_q(t)$ and $a_{q'}(t)$ need not be equal. If all corresponding centered coefficients are held fixed and only the clean log-odds is changed from $a_q(t)$ to $a_{q'}(t)$, permutation symmetry of the wrong classes gives zero zeroth-order contribution to equation 307. Likewise, the population updates of Q_e, Q_{z_μ}, Q_{z_e} are zero at zero query-side leakage for every clean log-odds because their query coordinates are centered.

Consequently, the effect of unequal clean log-odds contains one derivative with respect to the clean log-odds and one derivative with respect to either the wrong-score shape or query-side leakage. The second-derivative bounds proved in Appendix A.21, exact query balance, and equation 311–equation 312 give

$$\begin{aligned} &|\text{one-step population difference caused only by } a_q(t) \neq a_{q'}(t)| \\ &\quad \leq \frac{C\eta_{\mathrm{att}}\gamma^6}{M^2}(c_{\mathrm{wrong}} + c_{\mathrm{leak}}) \max_r p_r(t)(1-p_r(t))^2 |a_q(t) - a_{q'}(t)|. \end{aligned} \tag{317}$$

The early synchronization assumption gives

$$\left|\frac{1-p_q(t)}{p_q(t)} - \frac{1-p_{q'}(t)}{p_{q'}(t)}\right| \leq \frac{c_{\mathrm{sync}}}{2}.$$

Since

$$a_q(t) = -\log\frac{1-p_q(t)}{p_q(t)},$$

the mean-value theorem for the logarithm gives

$$|a_q(t) - a_{q'}(t)| \leq \frac{Cc_{\mathrm{sync}}}{\min\left\{\frac{1-p_q(t)}{p_q(t)}, \frac{1-p_{q'}(t)}{p_{q'}(t)}\right\}}. \tag{318}$$

The two inverse odds are comparable by a fixed factor on the stopped interval.

Clean progress gives

$$a_q(t+1) - a_q(t) \geq \frac{\eta_{\mathrm{att}}\gamma^6}{CM} \max_r p_r(t)(1-p_r(t))^2, \tag{319}$$

and the one-step upper bound gives $a_q(t+1) - a_q(t) \leq 1$. Therefore

$$\begin{aligned} &\frac{1-p_q(t)}{p_q(t)} - \frac{1-p_q(t+1)}{p_q(t+1)} \\ &\quad = \frac{1-p_q(t)}{p_q(t)} \left(1 - e^{-[a_q(t+1)-a_q(t)]}\right) \\ &\quad \geq \frac{\eta_{\mathrm{att}}\gamma^6}{CM} \frac{1-p_q(t)}{p_q(t)} \max_r p_r(t)(1-p_r(t))^2. \end{aligned} \tag{320}$$

Combining equation 317, equation 318, and equation 320, and summing until the $1/2$ crossing, gives

$$\begin{aligned} &\sum_t \frac{\eta_{\mathrm{att}}\gamma^6}{M^2} \max_r p_r(t)(1-p_r(t))^2 |a_q(t) - a_{q'}(t)| \\ &\quad \leq \frac{Cc_{\mathrm{sync}}}{M} \sum_t \frac{\dfrac{1-p_q(t)}{p_q(t)} - \dfrac{1-p_q(t+1)}{p_q(t+1)}}{\left(\dfrac{1-p_q(t)}{p_q(t)}\right)^2} \\ &\quad \leq \frac{Cc_{\mathrm{sync}}}{M}. \end{aligned} \tag{321}$$

The last inequality follows by comparison with

$$\int_1^\infty \frac{dx}{x^2} = 1.$$

Thus, including subsequent amplification by C_{prod}, the total contribution of unequal clean log-odds to any corresponding coefficient difference is at most

$$\frac{CC_{\mathrm{prod}}(c_{\mathrm{wrong}} + c_{\mathrm{leak}})c_{\mathrm{sync}}}{M}. \tag{322}$$

The fixed hierarchy of radii is chosen so that this is at most

$$\frac{c_{\mathrm{sync}}}{128(1+2c)M}.$$

We next control finite-context effects. The finite-context argument of Appendix A.21, applied to

$$h_s^{(n)}(t) + \frac{h_q^{(n)}(t)}{M-1}$$

in equation 307, gives the same one-step order as for each entry of Q_e, Q_{z_μ}, Q_{z_e}:

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{ML} \max_r p_r(t)(1-p_r(t))^2. \tag{323}$$

Hence the difference between the finite-context corrections in the two exchanged copies is bounded by the same expression up to a fixed factor. Using clean progress and summing until the $1/2$ crossing,

$$C_{\mathrm{prod}} \sum_t |\text{exchanged finite-context correction}| \leq \frac{CC_{\mathrm{prod}} \log(CM)}{L}. \tag{324}$$

Increasing only the numerical constant in the assumed lower bound on L makes this at most

$$\frac{c_{\mathrm{sync}}}{128(1+2c)M}.$$

It remains to control fresh-batch fluctuations. For every entry of Q_e, Q_{z_μ}, Q_{z_e}, Appendix A.21 gives accumulated conditional variance at most

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{BM}.$$

For the centered clean wrong-score coefficient in equation 306, the corresponding variance is no larger. The one-prompt Bernstein range is at most

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{B}.$$

The same bounds hold, up to a fixed factor, for the difference between the two exchanged copies.

Since subsequent population amplification is at most C_{prod}, the Bernstein–Ville inequality gives

$$\mathbb{P}\left(\sup_t \max_{\substack{q\neq q' \\ \text{centered clean wrong coefficients} \\ \text{entries of } Q_e, Q_{z_\mu}, Q_{z_e}}} |\text{centered fresh-batch exchanged contribution}| > \frac{c_{\mathrm{sync}}}{128(1+2c)M} \right)$$

$$\leq CM^2(M+M_z)^2 \exp\left[-\frac{c_{\mathrm{sync}}^2 B}{CC_{\mathrm{prod}}^2(1+2c)^2\eta_{\mathrm{att}}\gamma^6 M} \right]. \tag{325}$$

Under

$$B \geq CM \log \frac{C(M+M_z)^2 B(T+1)}{\delta},$$

with the numerical constant enlarged if necessary, the right-hand side of equation 325 is at most $\delta/16$.

Finally, the Phase-I handoff already bounds every corresponding exchanged coefficient at time T_1 by $c_{\mathrm{sync}}/(CM)$. Choosing its numerical constant sufficiently large, propagation by at most C_{prod} makes this contribution at most

$$\frac{c_{\mathrm{sync}}}{128(1+2c)M}.$$

Combining the propagated handoff contribution, the unequal-clean-log-odds contribution, the finite-context contribution, and the fresh-batch contribution gives, with probability at least $1-\delta/16$,

$$\begin{aligned} &|\text{centered clean wrong-score coefficient/entry of } Q_e, Q_{z_\mu}, Q_{z_e} \\ &- \text{its corresponding } q \leftrightarrow q' \text{ relabeling}| \leq \frac{c_{\mathrm{sync}}}{32(1+2c)M}. \end{aligned} \tag{326}$$

For a query with form index i, perturbation indices k_μ, k_e, and amplitudes $\kappa_{\mu,q}, \kappa_{e,q}$, the complete query-side preferred-versus-wrong contribution is

$$\begin{aligned} &(Q_e(t))_{q,i} - (Q_e(t))_{s,i} \\ &\quad + \kappa_{\mu,q}\left[(Q_{z_\mu}(t))_{q,k_\mu} - (Q_{z_\mu}(t))_{s,k_\mu} \right] \\ &\quad + \kappa_{e,q}\left[(Q_{z_e}(t))_{q,k_e} - (Q_{z_e}(t))_{s,k_e} \right]. \end{aligned} \tag{327}$$

Since $|\kappa_{\mu,q}|, |\kappa_{e,q}| \leq c$, equation 326 implies that the difference between equation 327 and its corresponding $q \leftrightarrow q'$ relabeling is at most

$$\frac{c_{\mathrm{sync}}}{16M}. \tag{328}$$

We next compare the conditional clean-log-odds drifts directly from the exact identity equation 163. This step is needed because the lower bound equation 164 alone does not determine the sign of the difference between two class drifts.

Fix $q \neq q'$ and couple their batches by the same label exchange as above. For the exactly exchanged, infinite-context reference profile with preferred mass $p_q(t)$ and equal wrong masses $(1-p_q(t))/(M-1)$, equation 148 gives

$$h_q^{(n)} = \frac{M}{M-1} p_q(t)(1-p_q(t))^2, \qquad h_s^{(n)} = -\frac{1}{M-1} h_q^{(n)} \quad (s \neq q). \tag{329}$$

Hence every clean preferred-versus-wrong gap has the same increment

$$\frac{\eta_{\mathrm{att}}\gamma^6}{M}\frac{M^2}{(M-1)^2}p_q(t)(1-p_q(t))^2. \tag{330}$$

Because all wrong-gap increments are identical, the logarithm in the exact formula equation 163 introduces no approximation: the clean log-odds increment is exactly the same common gap increment. The corresponding statement holds after exchanging q and q'.

For the actual stopped profiles, use the exact identity equation 163 for the two query classes and couple the two batches by the same $q \leftrightarrow q'$ exchange. The wrong-score tube gives, for every $s \neq q$,

$$\left|\frac{e^{A_{s,q}(t)}}{\sum_{r\neq q} e^{A_{r,q}(t)}} - \frac{1}{M-1}\right| \le \frac{Cc_{\mathrm{wrong}}}{M-1}, \tag{331}$$

and Lemma 6 gives, for every prompt,

$$\left|\pi_s^{(n)}(t) - \frac{1-p_n(t)}{M-1}\right| \le \frac{C(c_{\mathrm{wrong}}+c_{\mathrm{leak}}+\rho)}{M-1}(1-p_n(t)), \qquad s \neq j_n. \tag{332}$$

Insert equation 332 into the polynomial formula equation 148, and sum over the wrong classes before taking absolute values. Direct expansion gives

$$\left|h_{j_n}^{(n)} - \frac{M}{M-1}p_n(1-p_n)^2\right| \le C(c_{\mathrm{wrong}}+c_{\mathrm{leak}}+\rho)\,p_n(1-p_n)^2, \tag{333}$$

and, uniformly for $s \neq j_n$,

$$\left|h_{j_n}^{(n)} - h_s^{(n)} - \frac{M^2}{(M-1)^2}p_n(1-p_n)^2\right| \le C(c_{\mathrm{wrong}}+c_{\mathrm{leak}}+\rho)\,p_n(1-p_n)^2. \tag{334}$$

Under the exchanged coupling, the wrong-class deviations appearing in equation 331–equation 334 are the same when the corresponding probed coefficients are the same. Differentiate the remainder in equation 334 along the interpolation from $a_q(t)$ to $a_{q'}(t)$ while keeping those exchanged deviations coupled. Because $dp/da = p(1-p)$ and every term in the remainder contains the factor $p(1-p)^2$, the explicit degree-three polynomial and $dp/da = p(1-p)$ give

$$\left|\frac{d}{da}\left[h_{j_n}^{(n)} - h_s^{(n)} - \frac{M^2}{(M-1)^2}p(1-p)^2\right]\right| \le C(c_{\mathrm{wrong}}+c_{\mathrm{leak}}+\rho)\,p(1-p)^2. \tag{335}$$

The derivatives of the log-sum-exp weights in equation 163 have absolute row sum at most one. The mean-value theorem applied to equation 163 and equation 335 therefore shows that the common wrong-profile asymmetry contributes at most

$$\frac{C\eta_{\mathrm{att}}\gamma^6(c_{\mathrm{wrong}}+c_{\mathrm{leak}}+\rho)}{M}\max\{p_q(1-p_q)^2, p_{q'}(1-p_{q'})^2\}|a_q - a_{q'}|. \tag{336}$$

In the early regime, the only exchanged residual effects that must be controlled separately are the centered clean wrong-class shape and the three query-side leakage blocks. By equation 325 and equation 326, their contribution to the difference of the two conditional clean-log-odds increments is at most

$$\frac{\eta_{\mathrm{att}}\gamma^6 c_{\mathrm{sync}}}{CM^2}\left[p_q(t)(1-p_q(t))^2 + p_{q'}(t)(1-p_{q'}(t))^2\right]. \tag{337}$$

Indeed, the centered clean wrong-shape coefficient difference is $O(c_{\mathrm{sync}}/M)$ by equation 325, while the preferred-versus-wrong contribution of the three query-side blocks differs by $O(c_{\mathrm{sync}}/M)$ by equation 326. The summed first-derivative estimate equation 266, the fact that the derivatives of the log-sum-exp weights in equation 163 have absolute row sum at most one, and exact query balance supply the remaining factor $1/M$.

The context-side and two-sided blocks have class-independent infinite-context population contributions by equation 282; their finite-context effects enter through the correction below.

Finite-context difference between two query classes. We now compare the finite-context corrections for query classes q and q' directly. No additional parametrization of the wrong-class profile is needed.

Fix a wrong class $s \neq q$. Along the interpolation from the clean profile of query class q to that of query class q', keep the relative wrong-class proportions fixed and vary only the preferred mass. From the exact formula

$$h_a = \pi_a \left(\mathbf{1}\{a = q\} - p - \pi_a + \sum_r \pi_r^2 \right), \qquad p = \pi_q,$$

we obtain

$$\begin{aligned} h_q - h_s = p(1-p)^2 \left[1 + \sum_{r \neq q} \left(\frac{\pi_r}{1-p} \right)^2 \right] \\ + (1-p)^2 \frac{\pi_s}{1-p} \left[p + \frac{\pi_s}{1-p} - (1-p) \sum_{r \neq q} \left(\frac{\pi_r}{1-p} \right)^2 \right]. \end{aligned} \tag{338}$$

On the wrong-score tube,

$$\frac{\pi_r}{1-p} \leq \frac{C}{M}, \qquad r \neq q.$$

Differentiate equation 338 with respect to an additive class logit and then with respect to p, while keeping the relative wrong-class proportions fixed. Using

$$\frac{\partial p}{\partial z_q} = p(1-p), \qquad \frac{\partial p}{\partial z_k} = -p\pi_k, \quad k \neq q,$$

and

$$\sum_{k \neq q} \pi_k = 1 - p,$$

gives

$$\begin{aligned} \sum_b \left| \frac{\partial}{\partial p} \frac{\partial (h_q - h_s)}{\partial z_b} \right| + \sum_b \left| \frac{\partial}{\partial p} \frac{\partial^2 (h_q - h_s)}{\partial z_b^2} \right| \\ + \sum_{b,c} \left| \frac{\partial}{\partial p} \frac{\partial^2 (h_q - h_s)}{\partial z_b \partial z_c} \right| \leq C(1-p). \end{aligned} \tag{339}$$

Indeed, every term in equation 338 contains the factor $(1-p)^2$; differentiating once with respect to p removes at most one factor $1-p$. Every derivative with respect to a wrong-class logit carries a factor $\pi_r/(1-p)$, and these factors sum to one over the wrong classes. Hence summing over all logit indices produces no factor M.

We now apply exactly the same finite-context replacement argument as in Appendix A.21. For the random class counts, the first-order replacement term is centered and has conditional mean zero. The difference between the second-order remainders for query classes q and q' is, by equation 339, at most

$$\frac{CM}{L} |p_q(t) - p_{q'}(t)| \left[(1 - p_q(t)) + (1 - p_{q'}(t)) \right]. \tag{340}$$

The same estimate holds for the within-class nuisance averages, because their conditional first moments and second moments are respectively $O(M/L)$ and $O(M/L)$ and the same derivative bound equation 339 applies.

Exactly B/M prompts in the batch have each fixed query class. Therefore the attention update contributes the additional factor $\eta_{\mathrm{att}} \gamma^6 / M$, and the complete difference between the finite-context corrections for the two clean log-odds updates is bounded by

$$\frac{C \eta_{\mathrm{att}} \gamma^6}{L} |p_q(t) - p_{q'}(t)| \left[(1 - p_q(t)) + (1 - p_{q'}(t)) \right]. \tag{341}$$

We may now combine the equal-wrong reference calculation, the wrong-profile perturbation bound obtained above, the exchanged clean/query-side bound equation 328, and equation 341. This gives

$$\begin{aligned}
&\Big|\mathbb{E}_t[a_q(t+1) - a_q(t) - a_{q'}(t+1) + a_{q'}(t)] \\
&\qquad - \frac{\eta_{\mathrm{att}}\gamma^6}{M}\frac{M^2}{(M-1)^2}\Big[p_q(t)(1-p_q(t))^2 - p_{q'}(t)(1-p_{q'}(t))^2\Big]\Big| \\
&\le \frac{C\eta_{\mathrm{att}}\gamma^6(c_{\mathrm{wrong}} + c_{\mathrm{leak}} + \rho)}{M}\max\big\{p_q(t)(1-p_q(t))^2, p_{q'}(t)(1-p_{q'}(t))^2\big\}\,|a_q(t) - a_{q'}(t)| \\
&\quad + \frac{C\eta_{\mathrm{att}}\gamma^6}{L}|p_q(t) - p_{q'}(t)|\,[(1-p_q(t)) + (1-p_{q'}(t))] \\
&\quad + \frac{\eta_{\mathrm{att}}\gamma^6 c_{\mathrm{sync}}}{CM^2}\left[p_q(t)(1-p_q(t))^2 + p_{q'}(t)(1-p_{q'}(t))^2\right].
\end{aligned} \tag{342}$$

It remains only to verify that the first term on the right does not destroy the sign of the principal inverse-odds dynamics. This follows directly from the same explicit formula equation 338.

Keep the relative wrong-class proportions fixed and regard the preferred mass as

$$p = \frac{1}{1+e^{-a}}.$$

The wrong-score tube and the clean-to-prompt comparison imply that the difference between the actual preferred-versus-wrong coefficient and its equal-wrong value satisfies

$$\left|\frac{d}{d(e^{-a})}\left[e^{-a}\left(h_q - h_s - \frac{M^2}{(M-1)^2}p(1-p)^2\right)\right]\right| \le C(c_{\mathrm{wrong}} + c_{\mathrm{leak}} + \rho)\frac{e^{-2a}}{(1+e^{-a})^4}. \tag{343}$$

To see this, substitute $p = (1+e^{-a})^{-1}$ into equation 338. The equal-wrong terms cancel. Every remaining wrong-profile deviation is $O((c_{\mathrm{wrong}} + c_{\mathrm{leak}} + \rho)/M)$, while the sum over the $M-1$ wrong classes is bounded because the deviations are centered. Differentiating the remaining degree-three expression gives precisely the factor $e^{-2a}/(1+e^{-a})^4$.

For the equal-wrong term,

$$\frac{d}{d(e^{-a})}\left[e^{-a}\frac{M^2}{(M-1)^2}p(1-p)^2\right] = \frac{3M^2}{(M-1)^2}\frac{e^{-2a}}{(1+e^{-a})^4}. \tag{344}$$

Hence, after choosing the fixed radii so that

$$C(c_{\mathrm{wrong}} + c_{\mathrm{leak}} + \rho) < 1,$$

the perturbation in equation 343 is a strict fixed fraction of the positive derivative in equation 344. Therefore the complete infinite-context inverse-odds map remains nonexpansive throughout the early stopped interval.

We now close the stopped recursion separately before and after all clean preferred masses cross $1/2$.

Early regime: $\min_q p_q(t) \le 1/2$**.** Work directly with the inverse odds already appearing in equation 172. Since

$$\frac{1-p_q(t)}{p_q(t)} = e^{-a_q(t)}, \qquad e^{-a_q(t)}p_q(t)(1-p_q(t))^2 = (1-p_q(t))^3, \tag{345}$$

Taylor's theorem for the exponential, the one-step bound equation 342, and the signed drift estimate equation 342 give

$$\begin{aligned}
&\Big|\mathbb{E}_t\left[e^{-a_q(t+1)} - e^{-a_{q'}(t+1)}\right] - \left[e^{-a_q(t)} - e^{-a_{q'}(t)}\right] \\
&\qquad + \frac{\eta_{\mathrm{att}}\gamma^6}{M}\frac{M^2}{(M-1)^2}\left[\left(\frac{e^{-a_q(t)}}{1+e^{-a_q(t)}}\right)^3 - \left(\frac{e^{-a_{q'}(t)}}{1+e^{-a_{q'}(t)}}\right)^3\right]\Big| \\
&\le \frac{C\eta_{\mathrm{att}}\gamma^6}{L}\frac{\left|e^{-a_q(t)} - e^{-a_{q'}(t)}\right|}{1 + \min\{e^{-a_q(t)}, e^{-a_{q'}(t)}\}} \\
&\qquad + \frac{\eta_{\mathrm{att}}\gamma^6 c_{\mathrm{sync}}}{2^9 M^2}\left[\left(\frac{e^{-a_q(t)}}{1+e^{-a_q(t)}}\right)^3 + \left(\frac{e^{-a_{q'}(t)}}{1+e^{-a_{q'}(t)}}\right)^3\right].
\end{aligned} \tag{346}$$

The quadratic remainder from the exponential expansion causes no independent additive error: under the exchanged coupling it vanishes when the two exchanged states agree, and the mean-value theorem together with equation 166 bounds its difference by a fixed multiple of the two terms already displayed on the right-hand side of equation 346. It is therefore absorbed by the same numerical constants.

The principal term in equation 346 is signed. Indeed, for any $u, v > 0$,

$$\left(\frac{u}{1+u}\right)^3 - \left(\frac{v}{1+v}\right)^3 = (u-v)\int_0^1 \frac{3\,[v + s(u-v)]^2}{[1+v+s(u-v)]^4}\,ds. \tag{347}$$

The integral is nonnegative and at most $3/16$. Therefore, for sufficiently small $\eta_{\mathrm{att}}\gamma^6$, the first two lines on the left of equation 346 form a non-expansive deterministic map of the inverse-odds pair difference.

The finite-context term in equation 346 is multiplicative in that same pair difference. Comparing it with the principal decrease of each inverse odds and using the early clean progress from Lemma 7 gives the accumulated bound

$$\frac{CM}{L}\int_1^{CM} \frac{(1+u)^2}{u^3}\,du \le \frac{CM\log(CM)}{L}. \tag{348}$$

The numerical constant in equation 152 is chosen so that the exponential of this quantity is at most the remaining margin between current bound and the $c_{\mathrm{sync}}/2$ threshold in equation 172. The final term of equation 346 is a c_{sync}/M fraction of the principal inverse-odds decrease; summing it from the handoff to the $1/2$ crossing gives at most $c_{\mathrm{sync}}/16$ after fixing the numerical constant in equation 75.

It remains to control the centered current-batch contribution. We use a one-step stopped martingale for the inverse-odds difference. Conditional on the batches used before iteration t, the quantity

$$\begin{aligned} &e^{-a_q(t+1)} - e^{-a_{q'}(t+1)} \\ &\quad - \mathbb{E}_t\left[e^{-a_q(t+1)} - e^{-a_{q'}(t+1)}\right] \end{aligned} \tag{349}$$

has conditional mean zero. Exact query balance, equation 149, equation 168, and

$$e^{-a_q(t)}p_q(t)(1-p_q(t))^2 = (1-p_q(t))^3 \tag{350}$$

give total predictable quadratic variation, up to the $1/2$ crossing, at most

$$\frac{C\eta_{\mathrm{att}}\gamma^6 M}{B}, \tag{351}$$

and the one-prompt Bernstein range is $C\eta_{\mathrm{att}}\gamma^6/B$.

Stop a fixed pair q, q' at the earlier of the $1/2$ crossing and the first time its inverse-odds difference reaches $c_{\mathrm{sync}}/2$. By equation 343, equation 344, and the monotonicity calculation below, the common population part of the inverse-odds map is non-expansive on this stopped interval. The only deterministic terms that can increase the absolute pair difference are therefore the finite-context term in equation 346 and the early exchanged clean/query-side forcing.

For every $\tau > 0$ satisfying

$$\tau\,\frac{C\eta_{\mathrm{att}}\gamma^6}{B} < 3, \tag{352}$$

the following stopped process is a nonnegative supermartingale:

$$\exp\Bigg\{-\tau\sum_{s=T_1}^{t-1}\Bigg[\frac{c_{\mathrm{sync}}}{2}\frac{C\eta_{\mathrm{att}}\gamma^6}{L[1+\min\{e^{-a_q(s)},e^{-a_{q'}(s)}\}]} + \frac{\eta_{\mathrm{att}}\gamma^6 c_{\mathrm{sync}}}{CM^2}\left[\left(\frac{e^{-a_q(s)}}{1+e^{-a_q(s)}}\right)^3+\left(\frac{e^{-a_{q'}(s)}}{1+e^{-a_{q'}(s)}}\right)^3\right]\Bigg] - \sum_{s=T_1}^{t-1}\frac{\tau^2\mathbb{E}_s\left[\left(e^{-a_q(s+1)}-e^{-a_{q'}(s+1)} - \mathbb{E}_s[e^{-a_q(s+1)}-e^{-a_{q'}(s+1)}]\right)^2\right]}{2\left[1-\tau C\eta_{\mathrm{att}}\gamma^6/(3B)\right]}\Bigg\} \times\cosh\Big(\tau[e^{-a_q(t)}-e^{-a_{q'}(t)}]\Big). \tag{353}$$

To verify the one-step supermartingale inequality, condition on the batches used before iteration s. The common population part is non-expansive by equation 343, equation 344 and equation 347. On the stopped interval,

$$|e^{-a_q(s)}-e^{-a_{q'}(s)}|\le\frac{c_{\mathrm{sync}}}{2}, \tag{354}$$

so the finite-context multiplicative term costs at most the first term in the first sum in equation 353, while the early exchanged clean/query-side forcing costs at most the second term. Apply

$$\cosh(x+y)\le e^{|y|}\cosh(x) \tag{355}$$

to these predictable terms. The remaining current-batch part has conditional mean zero, range $C\eta_{\mathrm{att}}\gamma^6/B$, and the conditional variance used in equation 351; its conditional Bernstein bound is canceled by the second exponential sum in equation 353. This proves the supermartingale property.

The deterministic sum involving the finite-context term is bounded by equation 348, the deterministic early exchanged-score forcing is at most $c_{\mathrm{sync}}/16$, and the conditional-variance sum is bounded by equation 351. Optimizing τ and applying Ville's inequality gives a stopped fluctuation of at most $c_{\mathrm{sync}}/16$ for a fixed pair. A union bound over the at most M^2 pairs makes this simultaneous with failure probability at most $\delta/16$ under equation 152. Choosing these constants so that the inverse-odds difference at the last early iterate is in fact at most $c_{\mathrm{sync}}/4$; the lemma records the weaker bound $c_{\mathrm{sync}}/2$.

Late regime: direct synchronization from the clean-log-odds drift. Assume now that every $p_q(t)\ge 1/2$, and stop at the first failure of the leakage bound, the wrong-score tube, preferred maximality, KKT persistence, or

$$\max_{q,q'}|a_q(t)-a_{q'}(t)|\le c_{\mathrm{sync}}. \tag{356}$$

The late-regime argument works directly with the clean log-odds drift.

Fix a query class q. In the infinite-context calculation, the context-side and two-sided leakage blocks have no class-dependent effect, as proved in Appendix A.21. Thus only the three query-side blocks

$$Q_e(t),\qquad Q_{z_\mu}(t),\qquad Q_{z_e}(t)$$

can perturb the clean class logits.

For a query with form index i, perturbation indices k_μ,k_e, and perturbation amplitudes $\kappa_{\mu,q},\kappa_{e,q}$, the query-side contribution to the logit of main class a is

$$(Q_e(t))_{a,i}+\kappa_{\mu,q}(Q_{z_\mu}(t))_{a,k_\mu}+\kappa_{e,q}(Q_{z_e}(t))_{a,k_e}. \tag{357}$$

By the leakage tube,

$$\max_a\left|(Q_e(t))_{a,i}+\kappa_{\mu,q}(Q_{z_\mu}(t))_{a,k_\mu}+\kappa_{e,q}(Q_{z_e}(t))_{a,k_e}\right|\le Cc_{\mathrm{leak}}. \tag{358}$$

Each of the three query coordinates is centered:

$$\mathbb{E}\left[\mathbf{1}\{e_q=e_i\}-\frac{1}{M}\right]=0,$$

and

$$\mathbb{E}\left[\kappa_{\mu,q}\mathbf{1}\{z_{\mu,q}=z_{\mu,k}\}\right]=\mathbb{E}\left[\kappa_{e,q}\mathbf{1}\{z_{e,q}=z_{e,k}\}\right]=0.$$

Therefore the conditional expectation of equation 357 is zero.

Fix $s\neq q$. Regard h_q-h_s as a function of the additive class logits. Taylor expansion from zero query-side leakage to equation 357 has zero expected first-order term. The second-order term is controlled by equation 268 and equation 358. Hence

$$|\mathbb{E}_t[h_q-h_s]-(h_q-h_s)_{\mathrm{clean}}|\le Cc_{\mathrm{leak}}^2p_q(t)(1-p_q(t))^2. \tag{359}$$

Here $(h_q-h_s)_{\mathrm{clean}}$ means that the same quantity is evaluated using only the clean class logits $(A_{a,q}(t))_{a=1}^M$.

We next compare the actual clean wrong-class profile with the equal-wrong profile having the same preferred mass $p_q(t)$. The wrong-score tube gives, for every $s\neq q$,

$$\left|\frac{\pi_s(t)}{1-p_q(t)}-\frac{1}{M-1}\right|\le\frac{Cc_{\mathrm{wrong}}}{M-1}. \tag{360}$$

Because the normalized wrong masses sum to one,

$$\sum_{r\neq q}\left[\frac{\pi_r(t)}{1-p_q(t)}-\frac{1}{M-1}\right]=0. \tag{361}$$

Consequently,

$$\begin{aligned}&\sum_{r\neq q}\left(\frac{\pi_r(t)}{1-p_q(t)}\right)^2-\frac{1}{M-1}\\&\qquad=\sum_{r\neq q}\left[\frac{\pi_r(t)}{1-p_q(t)}-\frac{1}{M-1}\right]^2\le\frac{Cc_{\mathrm{wrong}}^2}{M-1}.\end{aligned} \tag{362}$$

From the exact definition of h_a,

$$h_q=p_q(t)(1-p_q(t))^2\left[1+\sum_{r\neq q}\left(\frac{\pi_r(t)}{1-p_q(t)}\right)^2\right], \tag{363}$$

$$\begin{aligned}h_s=&-(1-p_q(t))^2\frac{\pi_s(t)}{1-p_q(t)}\Bigg[p_q(t)+\frac{\pi_s(t)}{1-p_q(t)}\\&\qquad-(1-p_q(t))\sum_{r\neq q}\left(\frac{\pi_r(t)}{1-p_q(t)}\right)^2\Bigg].\end{aligned} \tag{364}$$

If all wrong masses are equal, these identities give exactly

$$h_q-h_s=\frac{M^2}{(M-1)^2}p_q(t)(1-p_q(t))^2. \tag{365}$$

Using equation 360–equation 362 in equation 363–equation 364 therefore gives

$$\left|(h_q-h_s)_{\mathrm{clean}}-\frac{M^2}{(M-1)^2}p_q(t)(1-p_q(t))^2\right|\le Cc_{\mathrm{wrong}}p_q(t)(1-p_q(t))^2. \tag{366}$$

Combining equation 359 and equation 366,

$$\left|\mathbb{E}_t[h_q-h_s]-\frac{M^2}{(M-1)^2}p_q(t)(1-p_q(t))^2\right|\le C(c_{\mathrm{wrong}}+c_{\mathrm{leak}}^2)p_q(t)(1-p_q(t))^2. \tag{367}$$

We now restore finite context. The finite-context bounds proved in Appendix A.21 give

$$\left|\mathbb{E}_th_q^{(n)}(t)-h_{q,\infty}^{(n)}(t)\right|\le\frac{CM}{L}p_q(t)(1-p_q(t))^2, \tag{368}$$

and, uniformly for $s \neq q$,

$$\left|\mathbb{E}_t h_s^{(n)}(t) - h_{s,\infty}^{(n)}(t)\right| \leq \frac{C}{L} p_q(t)(1-p_q(t))^2. \tag{369}$$

Exactly B/M prompts have query class q. Hence the finite-context correction to the conditional mean increment of every clean preferred-versus-wrong gap is at most

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{L} p_q(t)(1-p_q(t))^2. \tag{370}$$

The exact clean-log-odds identity in Lemma 7 expresses $a_q(t+1) - a_q(t)$ as a log-sum-exp of these clean gap increments. Preferred-class maximality gives

$$0 \leq S_{q,s}(t+1) - S_{q,s}(t) \leq \frac{C\eta_{\mathrm{att}}\gamma^6}{M} p_q(t)(1-p_q(t))^2 \leq 1. \tag{371}$$

For numbers in $[0,1]$, Taylor's theorem for the log-sum-exp gives

$$\begin{aligned}&\left|a_q(t+1) - a_q(t) - \sum_{s\neq q} \frac{e^{A_{s,q}(t)}}{\sum_{r\neq q} e^{A_{r,q}(t)}} \left[S_{q,s}(t+1) - S_{q,s}(t)\right]\right| \\ &\qquad \leq C \max_{s\neq q} \left[S_{q,s}(t+1) - S_{q,s}(t)\right]^2.\end{aligned} \tag{372}$$

By equation 371, the right side is at most

$$\frac{C\eta_{\mathrm{att}}^2\gamma^{12}}{M^2} p_q(t)^2(1-p_q(t))^4.$$

Using equation 367, equation 370, and equation 372, we conclude that

$$\begin{aligned}&\left|\mathbb{E}_t[a_q(t+1) - a_q(t)] - \frac{\eta_{\mathrm{att}}\gamma^6}{M}\frac{M^2}{(M-1)^2} p_q(t)(1-p_q(t))^2\right| \\ &\qquad \leq \frac{C\eta_{\mathrm{att}}\gamma^6}{M}\left(c_{\mathrm{wrong}} + c_{\mathrm{leak}}^2 + \frac{M}{L} + \frac{\eta_{\mathrm{att}}\gamma^6}{M}\right) p_q(t)(1-p_q(t))^2.\end{aligned} \tag{373}$$

Now fix $q \neq q'$. The principal scalar function is

$$p(1-p)^2. \tag{374}$$

For $p \geq 1/2$,

$$\frac{d}{da}\left[p(1-p)^2\right] = p(1-p)^2(1-3p) \leq -\frac{1}{2}p(1-p)^2. \tag{375}$$

Therefore, by the mean-value theorem and the stopped synchronization tube, the two class signals are comparable by a fixed factor and

$$\begin{aligned}&\operatorname{sgn}(a_q - a_{q'})\left[\frac{\eta_{\mathrm{att}}\gamma^6}{M}\frac{M^2}{(M-1)^2}\left(p_q(1-p_q)^2 - p_{q'}(1-p_{q'})^2\right)\right] \\ &\qquad \leq -\frac{\eta_{\mathrm{att}}\gamma^6}{CM}\min\{p_q(1-p_q)^2,\, p_{q'}(1-p_{q'})^2\}|a_q - a_{q'}|.\end{aligned} \tag{376}$$

Apply equation 373 to q and q' and use signal comparability. Whenever

$$|a_q(t) - a_{q'}(t)| \geq \frac{c_{\mathrm{sync}}}{2}, \tag{377}$$

equations equation 373–equation 376 give

$$\begin{aligned}&\operatorname{sgn}(a_q(t) - a_{q'}(t))\mathbb{E}_t\left[a_q(t+1) - a_{q'}(t+1) - a_q(t) + a_{q'}(t)\right] \\ &\leq -\frac{\eta_{\mathrm{att}}\gamma^6}{CM}\min\{p_q(t)(1-p_q(t))^2,\, p_{q'}(t)(1-p_{q'}(t))^2\}|a_q(t) - a_{q'}(t)|\end{aligned} \tag{378}$$

provided the already free numerical choices are made so that

$$C\left(c_{\mathrm{wrong}}+c_{\mathrm{leak}}^2+\frac{M}{L}+\frac{\eta_{\mathrm{att}}\gamma^6}{M}\right)\le\frac{c_{\mathrm{sync}}}{16}. \tag{379}$$

The fixed hierarchy equation 33, a sufficiently large constant in equation 152, and the existing small-step condition make equation 379 hold without changing the asymptotic requirements on L, B, or the step sizes.

Hence the conditional drift of every pair is strictly inward on the outer half of the late synchronization tube.

Fresh-batch fluctuations in the late regime. It remains to show that the fresh-batch randomness cannot push two clean log-odds from the interior of the synchronization tube to its boundary.

For fixed $q\neq q'$, exact query balance and the bounds on $h_a^{(n)}(t)$ give

$$\begin{aligned}&\mathbb{E}_t\Big[\big([a_q(t+1)-a_{q'}(t+1)]-\mathbb{E}_t[a_q(t+1)-a_{q'}(t+1)]\big)^2\Big]\\&\quad\le\frac{C\eta_{\mathrm{att}}^2\gamma^{12}}{BM}\Big[p_q(t)^2(1-p_q(t))^4+p_{q'}(t)^2(1-p_{q'}(t))^4\Big].\end{aligned} \tag{380}$$

Indeed, exactly B/M prompts have each fixed query class, each prompt contributes $O(\eta_{\mathrm{att}}\gamma^6/B)$ to the corresponding clean-gap update, and the squared class-gradient coefficients are bounded by a fixed multiple of $p_q(t)^2(1-p_q(t))^4$. The balanced cross-channel query assignment is handled by the same permutation-exposure argument used earlier for fresh batches.

Using synchronization and Lemma 8, equation 380 gives

$$\sum_t\mathbb{E}_t\Big[\big([a_q(t+1)-a_{q'}(t+1)]-\mathbb{E}_t[a_q(t+1)-a_{q'}(t+1)]\big)^2\Big]\le\frac{C\eta_{\mathrm{att}}\gamma^6}{B}. \tag{381}$$

Moreover, exposing the fresh batch one prompt or one permutation block at a time gives Bernstein range

$$\frac{C\eta_{\mathrm{att}}\gamma^6}{B}. \tag{382}$$

We first check the transition from the early regime to the late regime. At the last iteration before every preferred clean mass reaches $1/2$, the strengthened early estimate gives

$$\left|\frac{1-p_q(t)}{p_q(t)}-\frac{1-p_{q'}(t)}{p_{q'}(t)}\right|\le\frac{c_{\mathrm{sync}}}{4}\qquad\text{for every } q,q'. \tag{383}$$

At that iteration at least one query class still has preferred mass below $1/2$, and hence its inverse odds is larger than one. Therefore equation 383 implies that all inverse odds are bounded below by

$$1-\frac{c_{\mathrm{sync}}}{4}.$$

Choose the already small attention step size so that one update changes each inverse odds by at most $c_{\mathrm{sync}}/32$. At the first iteration at which every preferred mass is at least $1/2$, we then have

$$\left|\frac{1-p_q(t)}{p_q(t)}-\frac{1-p_{q'}(t)}{p_{q'}(t)}\right|\le\frac{5c_{\mathrm{sync}}}{16},\qquad\min_q\frac{1-p_q(t)}{p_q(t)}\ge\frac{3}{4}. \tag{384}$$

Since

$$a_q(t)=-\log\frac{1-p_q(t)}{p_q(t)},$$

the mean-value theorem and equation 384 give

$$\max_{q,q'}|a_q(t)-a_{q'}(t)|\le\frac{5c_{\mathrm{sync}}}{12}<\frac{c_{\mathrm{sync}}}{2}. \tag{385}$$

Thus the late regime begins strictly inside the inner half of the synchronization tube.

Now fix $q \neq q'$. By the restoring-drift estimate proved immediately above, whenever

$$\frac{c_{\mathrm{sync}}}{2} \leq |a_q(t) - a_{q'}(t)| < c_{\mathrm{sync}}, \tag{386}$$

the conditional mean increment points toward zero:

$$\operatorname{sgn}(a_q(t) - a_{q'}(t))\, \mathbb{E}_t \left[a_q(t+1) - a_{q'}(t+1) - a_q(t) + a_{q'}(t)\right] \leq 0. \tag{387}$$

Choose the same attention step size sufficiently small that one complete update changes $|a_q - a_{q'}|$ by at most $c_{\mathrm{sync}}/16$. Therefore, whenever a pair enters the interval equation 386 from below $c_{\mathrm{sync}}/2$, its entrance value is at most

$$\frac{9c_{\mathrm{sync}}}{16}. \tag{388}$$

Consider the evolution from such an entrance until the pair either returns below $c_{\mathrm{sync}}/2$ or reaches c_{sync}. Throughout this interval the conditional drift is inward by equation 387. Hence reaching c_{sync} before returning below $c_{\mathrm{sync}}/2$ requires the accumulated centered fresh-batch increments to have absolute value at least

$$c_{\mathrm{sync}} - \frac{9c_{\mathrm{sync}}}{16} = \frac{7c_{\mathrm{sync}}}{16} > \frac{c_{\mathrm{sync}}}{4}.$$

By the Bernstein–Freedman maximal inequality, equation 381, and equation 382, for any fixed pair and any fixed possible entrance iteration,

$$\begin{aligned}\mathbb{P}\left(\sup \left| \sum \Big([a_q(t+1) - a_{q'}(t+1)] - \mathbb{E}_t[a_q(t+1) - a_{q'}(t+1)] \Big) \right| \geq \frac{c_{\mathrm{sync}}}{4} \right) \\ \leq 2 \exp\left[- \frac{c_{\mathrm{sync}}^2}{C\left(\eta_{\mathrm{att}}\gamma^6/B + (\eta_{\mathrm{att}}\gamma^6/B)c_{\mathrm{sync}}\right)} \right].\end{aligned} \tag{389}$$

The supremum is taken from that entrance iteration until the pair leaves the interval equation 386.

There are at most M^2 ordered pairs and at most $T+1$ possible entrance iterations. Hence a union bound and the existing lower bound on B give

$$\mathbb{P}\left(\text{some pair reaches } c_{\mathrm{sync}} \text{ in the late regime}\right) \leq \frac{\delta}{16}, \tag{390}$$

after increasing only the numerical constant in the batch-size condition.

Therefore, with probability at least $1 - \delta/16$,

$$\max_{q,q'} |a_q(t) - a_{q'}(t)| < c_{\mathrm{sync}}$$

throughout the late regime.

Combining this with the early-regime estimate proves synchronization throughout the stopped Phase-II interval. The early and late fresh-batch events together fail with probability at most $\delta/8$. Hence the synchronization boundary cannot be the first boundary reached by the stopped process.

The head 2 argument is identical after exchanging the μ- and e-channels. This completes the proof of the synchronization lemma. □