TRANSFORMER PROVABLY HALLUCINATES ON HIERARCHICAL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can generate fluent and seemingly credible responses that nevertheless contain false statements or unsupported claims, a phenomenon commonly referred to as hallucination. In this paper, we theoretically investigate one particular form of hallucination through a hierarchical sequence classification problem, where a common signal token encodes a coarse label, and a rarer nuance token encodes a fine-grained label. We characterize hallucination as a model having high predicted probability mass on the correct coarse class but near-chance probability on the correct fine-grained label within that class. We prove that a shallow transformer trained by gradient flow exhibits this behavior when nuance tokens are sufficiently rare, and token noise is small. Our results identify a mechanism through which the interaction between data frequency and learned attention can make hallucination a prolonged intermediate phase of training: The model first learns the common signal, but attention weights to the signal tokens further suppress learning from the already rare nuance, producing a long training plateau where the model hallucinates on input sequences containing the fine-grained cue from nuance tokens. Experiments on an LLM-generated synthetic BirdText dataset exhibit analogous coarse-to-fine learning behavior.

1 INTRODUCTION

Large language models (LLMs) can generate fluent and seemingly credible responses that nevertheless contain false statements or claims unsupported by the available evidence, a phenomenon broadly referred to as hallucination (Ji et al., 2023; Huang et al., 2025). Empirical studies have documented such failures in summarization, question answering, and long-form biographical generation, demonstrating that coherent responses can combine accurate information with invented or unsupported details (Maynez et al., 2020; Lin et al., 2022; Min et al., 2023). These errors are not attributable to a single cause. Studies of factual learning show that recall depends strongly on the prevalence of relevant information in pretraining data and that knowledge gained through exposure can subsequently be forgotten (Kandpal et al., 2023; Chang et al., 2024). These observations motivate examining hallucination not only as a property of generated text, but also in relation to how models acquire, retain, and use factual knowledge.

One particular form of hallucination occurs when a model produces the right kind of answer but fails to identify the correct answer for the particular input. For example, a question about a country's capital may get a response of a valid capital city belonging to a different country: it respects the broad category while missing the subject-specific fact (Nichani et al., 2025). This distinction is consistent with empirical observations of uneven knowledge acquisition. Linguistic knowledge can develop earlier than factual knowledge during pretraining (Liu et al., 2021), while fine-tuning experiments exhibit intermediate stages in which answers are grammatically acceptable but remain factually incorrect (Gong et al., 2025). More directly, controlled studies of factual learning observe extended periods during which models capture generic attributes or relation-level patterns before learning the associations specific to individual entities (Nichani et al., 2025; Zucchet et al., 2025). These findings suggest that some hallucination behavior reflects a discriminative gap: a model may recognize what makes an answer plausible without reliably using the details that make it correct for the particular case.

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.

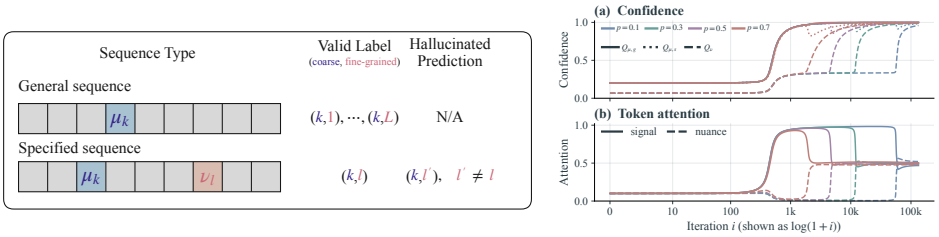


Figure 1: **Hallucination as an intermediate phase of coarse-to-fine learning.** We study a hierarchical classification problem in which every sequence contains a *signal token* μ_k identifying one of K coarse labels, whereas only a fraction p contains a *nuance token* ν_l identifying one of L fine-grained labels. The simplified transformer learns the coarse class early (solid and dotted lines) but remains uncertain about the fine-grained answer (dashed lines), even when the nuance token is present. Concentrated attention to the signal delays fine-grained learning, producing a prolonged hallucination plateau followed by a much shorter transition to learning the nuance. Smaller nuance token fraction p leads to longer plateaus. See later sections for details.

Motivated by these findings, we ask how the frequency of fine-grained information in the training dataset determines the duration of this discriminative gap. We study a hierarchical classification problem in which every sequence contains a *signal token* identifying one of K coarse classes, whereas only a fraction p contains a *nuance token* identifying one of L fine-grained labels. The remaining tokens carry no label information. Thus, *general sequences* reveal only the coarse label, while *specified sequences* also contain the cue needed to determine the full label.

We analyze a simplified, single-head transformer with softmax attention. We show that gradient flow on the population cross-entropy loss admits an exact reduction to a system of four-scalar ordinary differential equations. These dynamics reveal how learning the coarse label can delay fine-grained learning: because coarse information is available in every sequence, attention concentrates early on signal tokens and suppresses the contribution of nuance tokens even in specified sequences. The resulting delay therefore reflects both the rarity of fine-grained information and the evolution of attention, rather than rarity alone. Our contributions are as follows:

- **Reduction of gradient flow to four scalar ODEs.** We show that population gradient flow for the simplified transformer admits an exact reduction to four scalar ordinary differential equations under uniform label prior, orthonormal token patterns, and zero initialization (Section 3.1). The reduced system captures the coupled evolution of the attention query and prediction head, enabling a direct analysis of how the model learns coarse and fine-grained information.
- **Provable hallucination in the transformer model.** For sufficiently rare nuance tokens and suitably small token noise, we prove that training exhibits a plateau of hallucination phase: the expected probability of the correct coarse class is close to one, while that of the correct full label remains near the within-class random guessing even on inputs containing a nuance token (Theorem 2). We quantify the duration of this plateau and the much shorter subsequent transition to fine-grained learning, showing how attention to the common signal suppresses learning of the rarer nuance.
- **Experiments on the synthetic BirdText dataset.** We construct BirdText, a synthetic dataset of bird descriptions with controlled coarse and fine-grained attributes. Experiments with a Transformer classifier using frozen BERT features and a LoRA-adapted GPT-2 exhibit analogous coarse-to-fine learning behavior, including a prolonged species-generation plateau when descriptions containing fine-grained cues are rare (Figure 4).

Together, these results identify a mechanism by which data frequency and learned attention jointly lead to hallucination, stagnating fine-grained learning even when the necessary cue is present in the input. We provide an extended discussion on the related work in Appendix A.

2 Problem Settings

Hierarchical classification problem. Consider the sequence $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T] \in \mathbb{R}^{d\times T}$ of $T \geq 2$ tokens of dimension d and the multi-indexed label $\boldsymbol{y} = (k, l) \in [K] \times [L]$, where we call k the *coarse label* and l the *fine-grained label*. A data point $(\boldsymbol{X}, \boldsymbol{y})$ is sampled from a

distribution $\mathcal{D}(\mathcal{V}, p, \sigma)$ that depends on the following quantities: a set of d-dimensional orthonormal vectors $\mathcal{V} = \{\boldsymbol{\mu}_k\}_{k=1}^K \cup \{\boldsymbol{\nu}_l\}_{l=1}^L \cup \{\boldsymbol{\xi}_m\}_{m=1}^M$ (i.e. $\|\boldsymbol{v}\| = 1$ and $\langle \boldsymbol{v}, \boldsymbol{v}' \rangle = 0, \forall \boldsymbol{v} \neq \boldsymbol{v}' \in \mathcal{V}$) with $K, L \geq 2, M \geq 1$ and $K + L + M \leq d$, where we call $\{\boldsymbol{\mu}_k\}_{k=1}^K, \{\boldsymbol{\nu}_l\}_{l=1}^L, \{\boldsymbol{\xi}_m\}_{m=1}^M$ the *signal*, *nuance*, *non-relevant patterns* respectively; a *nuance probability* $0 \leq p \leq 1$; a noise level $\sigma \geq 0$. The sampling procedure can be described as:

1. Sample $\boldsymbol{y} \sim \mathrm{Unif}([K] \times [L])$;
2. Sample $t_s \sim \mathrm{Unif}([T])$. Let $\boldsymbol{x}_{t_s}^{\mathrm{clean}} = \boldsymbol{\mu}_k$;
3. Sample $z \sim \mathrm{Ber}(p)$. If $z = 0$, $\boldsymbol{x}_t^{\mathrm{clean}} \sim \mathrm{Unif}(\{\boldsymbol{\xi}_m\}_{m=1}^M), \forall t \in [T] \setminus \{t_s\}$; If $z = 1$, sample $t_n \sim \mathrm{Unif}([T] \setminus \{t_s\})$, let $\boldsymbol{x}_{t_n}^{\mathrm{clean}} = \boldsymbol{\nu}_l$ and $\boldsymbol{x}_t^{\mathrm{clean}} \sim \mathrm{Unif}(\{\boldsymbol{\xi}_m\}_{m=1}^M), \forall t \in [T] \setminus \{t_s, t_n\}$;
4. $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T]$, where $\boldsymbol{x}_t = \boldsymbol{x}_t^{\mathrm{clean}} + \sigma \boldsymbol{\varepsilon}_t / \sqrt{d}$, $\boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Roughly speaking, after the label $\boldsymbol{y}$ is determined, the Bernoulli random variable z determines how much the label information is revealed in the sampled $\boldsymbol{X}$ sequence. When $z = 0$, $\boldsymbol{X}$ is called a *general sequence* and has exactly one of its tokens (we call it a *signal token*), at a uniformly random position t_s, having the signal pattern $\boldsymbol{\mu}_k$ associated with the coarse label k, while the rest of the tokens have non-relevant patterns; Critically, observing a general sequence $\boldsymbol{X}$ alone cannot fully recover the label $\boldsymbol{y}$. When $z = 1$, $\boldsymbol{X}$ is called a *specified sequence* and has exactly one additional token (we call it a *nuance token*), at a position t_n, having the nuance pattern $\boldsymbol{\nu}_l$ associated with the fine-grained label l; this added nuance token allows one to infer the full label $\boldsymbol{y}$ with high accuracy from a specified sequence. Our analysis will also consider the conditional distributions $\mathcal{D}_g$, $\mathcal{D}_s$ of $\mathcal{D}$ on the events $z = 0$ and $z = 1$ respectively. Then the original $\mathcal{D}$ can be viewed as a mixture model $(1 - p)\mathcal{D}_g + p\mathcal{D}_s$, where the mixing probability is determined by the nuance probability p.

Evaluating classification models. With the construction of $(\boldsymbol{X}, \boldsymbol{y})$ above, a classification model produces a prediction $\hat{\boldsymbol{y}} = (\hat{k}, \hat{l})$ of the true label $\boldsymbol{y}$ from an observation $\boldsymbol{X}$. No model can achieve near-perfect prediction accuracy for general sequences $\boldsymbol{X}$ due to the missing fine-grained label information. Therefore, a classification model is good if it can correctly predict the coarse label (i.e., $\hat{k} = k$) for general sequences, and predict the full label (i.e., $\hat{\boldsymbol{y}} = \boldsymbol{y}$) for specified sequences.

We focus on a particular failure mode: the model assigns high probability mass to the correct coarse class but does not concentrate probability on the correct full label, even when the input contains a nuance cue. We use this coarse–fine discrimination gap as a stylized, probability-based notion of hallucination. Given a classification model $\boldsymbol{f} : \mathbb{R}^{d \times T} \to \mathbb{R}^{(KL)}$ that makes prediction on $\boldsymbol{y}$ from observed $\boldsymbol{X}$ through the probability distribution $\boldsymbol{q} = \left[q_{(1,1)}, \cdots, q_{(k,l)}, \cdots, q_{(K,L)}\right] = \mathrm{Softmax}(\boldsymbol{f}(\boldsymbol{X}))$, define

$$Q_{\mu,g}(\boldsymbol{f}) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{y}) \sim \mathcal{D}_g} \sum_{l=1}^L \boldsymbol{q}_{(k(\boldsymbol{y}),l)}\,, \quad Q_{\mu,s}(\boldsymbol{f}) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{y}) \sim \mathcal{D}_s} \sum_{l=1}^L \boldsymbol{q}_{(k(\boldsymbol{y}),l)}\,, \tag{1}$$

$$Q_{\nu}(\boldsymbol{f}) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{y}) \sim \mathcal{D}_s} \boldsymbol{q}_{\boldsymbol{y}}\,, \tag{2}$$

where we conveniently let $k(\boldsymbol{y})$ represent the coarse label in $\boldsymbol{y}$. When a prediction $\hat{\boldsymbol{y}}$ is sampled from the probability distribution $\boldsymbol{q}$, $Q_{\mu,g}$ and $Q_{\mu,s}$ represent the prediction accuracy of the coarse label k for the conditioned general and specified distributions $\mathcal{D}_g$ and $\mathcal{D}_s$, respectively; and Q_ν represents the prediction accuracy of the full label $\boldsymbol{y}$ for the conditioned specified sequence distribution $\mathcal{D}_s$. With this, we formally define our evaluation metrics for our theoretical studies:

Definition 1 (Hallucination on hierarchical classification)**.** *Given a tolerance $\varepsilon > 0$, a model $\boldsymbol{f}$ **hallucinates** if $Q_{\mu,g}(\boldsymbol{f}) \geq 1 - \varepsilon$, $Q_{\mu,s}(\boldsymbol{f}) \geq 1 - \varepsilon$ and $|Q_\nu(\boldsymbol{f}) - 1/L| \leq \varepsilon$.*

Remark 1. *Definition 1 characterizes hallucination by the probability masses of the model prediction. However, this metric can fail to evaluate the model prediction under argmax (in classification models) and greedy sampling (in LLMs), since the greedy prediction can succeed as long as there is a positive margin, however small it is, between the prediction probability of the correct label and the rest. But one can easily see that probability masses are better-behaved objects to analyze, which have led us to the current definitions. Nonetheless, we have reported the argmax/greedy prediction in all the numerical experiments we have considered, and hallucination persists.*

Training transformers. Consider the following simplified shallow transformer model $\boldsymbol{f} : \mathbb{R}^{d \times T} \times \mathbb{R}^{(KL) \times d} \times \mathbb{R}^d \to \mathbb{R}^{(KL)}$, formally defined as

$$\boldsymbol{f}_{\boldsymbol{W},\boldsymbol{v}}(\boldsymbol{X}) = \boldsymbol{W}\boldsymbol{X}\mathrm{Softmax}\left(\boldsymbol{X}^\top \boldsymbol{v}\right)\,. \tag{3}$$

We call $\boldsymbol{W}$ the *prediction head* and $\boldsymbol{v}$ the *query vector*. Given a data $(\boldsymbol{X}, \boldsymbol{y})$, the cross-entropy loss

$$\ell_{\text{CE}}(\boldsymbol{y}, \boldsymbol{f}_{\boldsymbol{W},\boldsymbol{v}}(\boldsymbol{X})) = -\log q_{\boldsymbol{y}}, \quad q_{\boldsymbol{y}} = \exp([\boldsymbol{f}_{\boldsymbol{W},\boldsymbol{v}}(\boldsymbol{X})]_{\boldsymbol{y}}) / \sum_{k=1,l=1}^{K,L} \exp([\boldsymbol{f}_{\boldsymbol{W},\boldsymbol{v}}(\boldsymbol{X})]_{(k,l)}) \tag{4}$$

evaluates how well the probability measure $\boldsymbol{q} = \text{Softmax}(\boldsymbol{f}_{\boldsymbol{W},\boldsymbol{v}}(\boldsymbol{X}))$ can predict the true label $\boldsymbol{y}$. One can then define a population loss

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{v}) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{y})\sim\mathcal{D}(\mathcal{V},p,\sigma)} \left(\ell_{\text{CE}}(\boldsymbol{y}, \boldsymbol{f}_{\boldsymbol{W},\boldsymbol{v}}(\boldsymbol{X}))\right) \tag{5}$$

and train the transformer model through the following gradient flow dynamics:

$$\dot{\boldsymbol{W}} = -\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}, \boldsymbol{v}),\ \dot{\boldsymbol{v}} = -\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{W}, \boldsymbol{v}), \quad \boldsymbol{W}(0) = \boldsymbol{0}, \boldsymbol{v}(0) = \boldsymbol{0}\,. \tag{6}$$

Our study mainly concerns the solution $\{\boldsymbol{W}(t), \boldsymbol{v}(t)\}$ of (6). As shown in the preview in the introduction, when the data distribution has a low nuance probability p, there exists a long period of time when the model $\boldsymbol{f}_{\boldsymbol{W}(t),\boldsymbol{v}(t)}$ hallucinates according to Definition 1. Our theoretical results precisely characterize such observations; We shall discuss them next.

3 Main Results

In this section, we present our main theoretical analysis for (6). First, we show that the gradient flow dynamics (6) can be equivalently described by a four-scalar nonlinear ODE system, where each scalar represents the projection of the prediction head $\boldsymbol{W}$ or the query vector $\boldsymbol{v}$ onto a one-dimensional subspace associated with the signal or the nuance tokens. Then, our main theorem shows that when the nuance probability p is sufficiently small, the model hallucinates because of sequential learning along the gradient flow: signal patterns are learned much earlier than nuance patterns, and the learning-time gap increases as p decreases.

3.1 Reduced ODE Systems

The reduction of the dynamics $\{\boldsymbol{W}(t), \boldsymbol{v}(t)\}$ to a four-scalar system is conceptually simple, in the noiseless ($\sigma = 0$) case: First of all, it is easy to see that the row space of $\boldsymbol{W}$ and the query vector $\boldsymbol{v}$ are restricted within the span of $\mathcal{V}$, by simply inspecting the gradients $\nabla_{\boldsymbol{W}}\mathcal{L}, \nabla_{\boldsymbol{v}}\mathcal{L}$. Moreover, due to the permutation symmetry of each pattern group (i.e., the flow (6) remains the same if one swaps two signal patterns, and the same for nuance or non-relevant patterns), the query vector $\boldsymbol{v}$ (we discuss $\boldsymbol{W}$ later) has identical projection on every signal (nuance, non-relevant) pattern; It is natural to consider

$$\boldsymbol{v}(t) = \bar{v}_\mu(t)\boldsymbol{\mu} + \bar{v}_\nu(t)\boldsymbol{\nu} + \bar{v}_\xi(t)\boldsymbol{\xi}, \quad \text{where } \boldsymbol{\mu} = \textstyle\sum_{k=1}^{K} \boldsymbol{\mu}_k, \boldsymbol{\nu} = \sum_{l=1}^{L} \boldsymbol{\nu}_l, \boldsymbol{\xi} = \sum_{m=1}^{M} \boldsymbol{\xi}_m\,. \tag{7}$$

Finally, we have the shift symmetry of $\boldsymbol{v}$: any perturbation of $\boldsymbol{v}$ along the direction of $\boldsymbol{\mu} + \boldsymbol{\nu} + \boldsymbol{\xi}$ induces a uniform attention-score shift that does not affect the softmax attention output, so the loss remains the same, i.e., $\mathcal{L}(\boldsymbol{W}, \boldsymbol{v}) = \mathcal{L}(\boldsymbol{W}, \boldsymbol{v} + c(\boldsymbol{\mu} + \boldsymbol{\nu} + \boldsymbol{\xi})), \forall c \in \mathbb{R}$; such shift symmetry[1] leads to an invariance that $K\bar{v}_\mu(t) + L\bar{v}_\nu(t) + M\bar{v}_\xi(t) \equiv 0, \forall t$, removing another degree of freedom for $\boldsymbol{v}(t)$. Therefore, the full $\boldsymbol{v}(t)$ can be characterized by the following two quantities:

$$v_\mu(t) = \bar{v}_\mu(t) - \bar{v}_\xi(t), \quad v_\nu(t) = \bar{v}_\nu(t) - \bar{v}_\xi(t)\,. \tag{8}$$

Similarly, using the same permutation symmetry and another shift symmetry of $\boldsymbol{W}$ that leaves the softmax output probability invariant, we can write

$$\boldsymbol{W}(t) = w_\mu(t)\boldsymbol{S} + w_\nu(t)\boldsymbol{N}, \text{ where } \boldsymbol{S} = \begin{bmatrix} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}/K)^\top \\ \vdots \\ (\boldsymbol{\mu}_K - \boldsymbol{\mu}/K)^\top \end{bmatrix} \otimes \mathbb{1}_L, \boldsymbol{N} = \mathbb{1}_K \otimes \begin{bmatrix} (\boldsymbol{\nu}_1 - \boldsymbol{\nu}/L)^\top \\ \vdots \\ (\boldsymbol{\nu}_L - \boldsymbol{\nu}/L)^\top \end{bmatrix}, \tag{9}$$

where $\otimes$ denotes the Kronecker product. Therefore, $\boldsymbol{W}(t), \boldsymbol{v}(t)$ can be fully described by the dynamics of these four scalars $v_\mu(t), v_\nu(t), w_\mu(t), w_\nu(t)$. We have been motivating such a reduction in the noiseless case, but it holds generally for any $\sigma \geq 0$, as suggested by the following theorem:

[1]This symmetry breaks when $\sigma > 0$. However, a weaker reflective symmetry holds under noise, which suffices for the same ODE reduction. Please see Appendix E.2 for details.

Theorem 1 (Reduced ODE dynamics)**.** *Consider the reduced population loss*

$$\mathcal{L}_s(w_\mu, w_\nu, v_\mu, v_\nu) = \mathcal{L}(\boldsymbol{W}(w_\mu, w_\nu), \boldsymbol{v}(v_\mu, v_\nu)), \tag{10}$$

where

$$\boldsymbol{W}(w_\mu, w_\nu) = w_\mu \boldsymbol{S} + w_\nu \boldsymbol{N}, \quad \boldsymbol{v}(v_\mu, v_\nu) = v_\mu \boldsymbol{\mu} + v_\nu \boldsymbol{\nu} - \tfrac{K v_\mu + L v_\nu}{K+L+M}(\boldsymbol{\mu} + \boldsymbol{\nu} + \boldsymbol{\xi}), \tag{11}$$

and the following four-scalar ODE system:

$$\begin{aligned} \begin{bmatrix} \dot{w_\mu} \\ \dot{w_\nu} \\ \dot{v_\mu} \\ \dot{v_\nu} \end{bmatrix} &= - \begin{bmatrix} 1/L(K-1) & 0 & 0 & 0 \\ 0 & 1/K(L-1) & 0 & 0 \\ 0 & 0 & 1/K + 1/M & 1/M \\ 0 & 0 & 1/M & 1/L + 1/M \end{bmatrix} \nabla_{(w_\mu, w_\nu, v_\mu, v_\nu)} \mathcal{L}_s \,, \\ w_\mu(0) = w_\nu(0) &= v_\mu(0) = v_\nu(0) = 0 \,. \end{aligned} \tag{12}$$

The solution $v_\mu(t), v_\nu(t), w_\mu(t), w_\nu(t)$ of (12) *recovers the solution $\boldsymbol{W}(t), \boldsymbol{v}(t)$ of* (6) *through* (11)*.*

Interpretations of the scalars. A four-scalar dynamics not only reduces the scale of the problem at hand, but also offers a clean interpretation of the transformer model during training. The query vector $\boldsymbol{v}$ is determined by two scalars v_μ and v_ν, whose scales affect how much attention scores the signal token and the nuance token (if present) receive. Ideally, both should be positive to filter out the non-relevant tokens from the output of the attention module $\boldsymbol{X}\mathrm{Softmax}\left(\boldsymbol{X}^\top \boldsymbol{v}\right)$, but their relative scale should not be too large; for example, when $v_\mu \gg v_\nu$, the presence of the nuance token diminishes due to the small attention weight it receives. The two-scalar representation of the prediction head $\boldsymbol{W}$ is even simpler: the scale of w_μ (or w_ν) exactly characterizes how confidently the head classifies the coarse (or the fine-grained) label when provided a signal (or a nuance) token; when a mixture of signal and nuance token is present, the head combines the predictions (in terms of prediction logits) of the two labels into a prediction on the full label.

With these interpretations, analyzing the growth and the relative scale of these four scalars through (12) allows us to directly probe into the prediction accuracy of the transformer model on general and specified sequences during training. We shall discuss our main results on hallucination next.

3.2 Transformer Provably Hallucinates During Training

Analyzing (12) in the case of small nuance probability and token noise leads to our main results:

Theorem 2 (Hallucination due to sequential learning)**.** *Let $\gamma = \frac{2M}{K+M}$. Given $0 < \varepsilon < (1 - 1/L)/2$, $\exists p_\varepsilon \in (0,1)$, $\tau_\varepsilon > 0$, $0 < c_\varepsilon < 1, C_\varepsilon > 1$, depending only on $(K, L, M, T, \varepsilon)$, such that whenever*

$$0 < p \le p_\varepsilon, \qquad 0 \le \tfrac{\sigma^2}{d} \le \tfrac{p}{\left(\log \frac{1}{p}\right)^\gamma}, \tag{13}$$

the solution $\boldsymbol{W}(t), \boldsymbol{v}(t)$ of (6) *satisfies the following: there exist times $\tau_\varepsilon < \tau_{\nu,0} < \tau_{\nu,1} < \infty$ with*

$$c_\varepsilon \tfrac{\left(\log \frac{1}{p}\right)^\gamma}{p} \le \tau_{\nu,0} \le C_\varepsilon \tfrac{\left(\log \frac{1}{p}\right)^\gamma}{p}, \qquad \tfrac{c_\varepsilon}{p} \le \tau_{\nu,1} - \tau_{\nu,0} \le \tfrac{C_\varepsilon}{p}, \tag{14}$$

such that

(i) The transformer model $\boldsymbol{f}_{\boldsymbol{W}(t), \boldsymbol{v}(t)} := \boldsymbol{f}_t$ hallucinates during $[\tau_\varepsilon, \tau_{\nu,0}]$:

$$\forall t \in [\tau_\varepsilon, \tau_{\nu,0}], \quad Q_{\mu,g}(\boldsymbol{f}_t), Q_{\mu,s}(\boldsymbol{f}_t) \ge 1 - \varepsilon, \quad |Q_\nu(\boldsymbol{f}_t) - 1/L| \le \varepsilon \,; \tag{15}$$

(ii) The transformer model $\boldsymbol{f}_t$ eventually solves the classification problem after $\tau_{\nu,1}$:

$$\forall t \ge \tau_{\nu,1}, \quad Q_{\mu,g}(\boldsymbol{f}_t), Q_{\mu,s}(\boldsymbol{f}_t), Q_\nu(\boldsymbol{f}_t) \ge 1 - \varepsilon \,. \tag{16}$$

We make the following remarks:

Hallucination in hierarchical classifications. From Theorem 2, we see that the training of the simplified transformer can be split into three phases: during the *signal-learning phase* ($t \le \tau_\varepsilon$), the model learns to extract the signal token from the input sequences and classify the coarse label with good accuracy, for both general and specified sequences. Then the training experiences a long

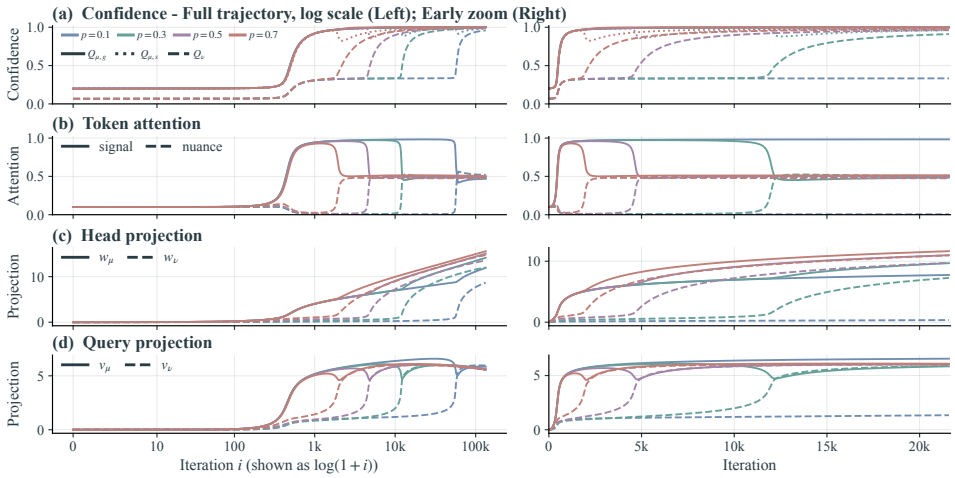


Figure 2: Training a simplified transformer on a hierarchical classification dataset (Experiment details in Appendix B.1). During the early training phase, the signal-subspace components w_μ, v_μ of the prediction head and query grow fast, irrespective of the nuance probability p, as shown in (c)(d), which enable the model to confidently predict the coarse label of both general and specified sequences. However, the nuance component v_ν grows much slower, resulting in a large discrepancy in scale when compared to its signal counterpart; the attention weights therefore concentrate on the signal token (See (b)), further stagnating the learning of nuance information. This eventually leads to a long plateau phase of hallucination, as shown in (a); and the smaller p, the longer the plateau.

plateau of the *hallucination phase* ($t \in [\tau_\varepsilon, \tau_{\nu,0}]$), whose length scales as $\Theta\left(\left(\log \frac{1}{p}\right)^\gamma / p\right)$, since it fails to extract the nuance token in the specified sequences through softmax attention (the actual proof makes this point clear, and it is also shown in our numerical validation in Figure 2). Finally, there is a nuance-learning phase ($t \in [\tau_{\nu,0}, \tau_{\nu,1}]$), after which the transformer model produces good predictions on both general and specified sequences.

It is tempting to attribute the hallucination results above to the small nuance probability p: the specified sequence takes a small portion of the data distribution; thus the model learns the nuance patterns much more slowly than the signal patterns. Indeed, this slowing-down effect is clearly seen from the fact that the length $\tau_{\nu,1} - \tau_{\nu,0}$ of the nuance-learning phase scales as $\Theta\left(1/p\right)$. However, it does not explain the full three-phase learning. Our analysis suggests the following:

Explanations of the three-phase learning. First of all, the hierarchical label and data structure are also important. For solely predicting the coarse label, the specified sequence can be regarded as general sequences as well; thus, the model always learns to predict the correct coarse label after a p-independent signal-learning phase. Moreover, the model structure and the associated training dynamics also play a role here. In particular, after the signal learning phase, the attention weights have concentrated mostly on the signal token due to the rapid growth of query vector $\boldsymbol{v}$ along the signal direction $\boldsymbol{\mu}$ even for specified sequences, which further suppresses the presence of the nuance token in $\nabla_{\boldsymbol{v}}\mathcal{L}$ on top of the effect the small p already has. The joint stagnating effects create a plateau phase of hallucination that asymptotically has longer duration $\Theta\left(\left(\log \frac{1}{p}\right)^\gamma / p\right)$ than the nuance-learning phase $\Theta\left(1/p\right)$. While the result itself requires a sufficiently small p, the long hallucination phase is still prominent for moderately small p in the numerical validation, as shown in Figure 2.

Comparison with Nichani et al. (2025). In terms of problem formulation and the analysis, the closest to ours is Nichani et al. (2025), where their data model is essentially the specified sequence model in this paper, and they also study the same hallucination behavior. However, there are several differences between our work and Nichani et al. (2025): First of all, we study the case when the nuance information might not be available in a new type of general sequences, whereas Nichani et al. (2025) concerns the case when the number of fine-grained classes is much more than the coarse ones, i.e., $L \gg K$ (stated in our terminology); therefore, we study two different factors that may cause the model to hallucinate. Secondly, the technical tools are different. Our analysis relies on a reduction to a small ODE, and Nichani et al. (2025) analyzes the weight evolutions directly; through the ODE analysis, we have shown an explicit dependence of the length of the hallucination plateau on the nuance probability p, while Nichani et al. (2025) have only shown the existence of a time epoch

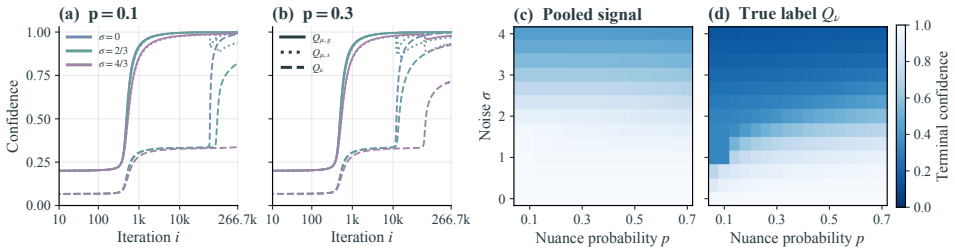


Figure 3: Training a simplified transformer on a hierarchical classification dataset with varying noise levels (Experiment details in Appendix B.1). The trained transformer can predict the coarse label confidently regardless of p as long as the noise level is not too high, as indicated by its terminal pooled signal confidence $(1-p)Q_{\mu,g} + pQ_{\mu,s}$ in (c). However, in the small p regime, the noise level should be lower so that the trained model can predict the fine-grained level as well, as shown in (d).

when hallucination occurs. Overall, our work and Nichani et al. (2025) have investigated model hallucination from complementary perspectives, and we believe practical hallucination should be better understood as a joint effect of both, plus extra factors that have not been theoretically explored.

Intermediate noise prevents nuance learning. We also present a complementary result where the noise level σ^2 is above the threshold prescribed by Theorem 2 but not too much.

Theorem 3 (Intermediate noise regime)**.** *Given $0 < \varepsilon < (1-1/L)/2$, $\exists p_\varepsilon \in (0,1)$, $\tau_\varepsilon > 0$, $C_\varepsilon > 1$, depending only on $(K, L, M, T, \varepsilon)$, such that whenever*

$$0 < p \le p_\varepsilon, \quad C_\varepsilon p \le \frac{\sigma^2}{d} \le C_\varepsilon p_\varepsilon, \tag{17}$$

then the transformer model $\boldsymbol{f}_t$ satisfies that:

$$Q_{\mu,g}(\boldsymbol{f}_t) \ge 1-\varepsilon, \ \forall t \ge \tau_\varepsilon, \qquad 0 \le Q_\nu(\boldsymbol{f}_t) - \frac{Q_{\mu,s}(\boldsymbol{f}_t)}{L} \le \varepsilon, \ \forall t \ge 0. \tag{18}$$

In this case, the signal learning succeeds on general sequences while the theorem provides no guarantee for coarse-label prediction for specified sequences. Nonetheless, the nuance learning can never be achieved, from the bound that $Q_\nu(\boldsymbol{f}_t) \simeq {}^{Q_{\mu,s}(\boldsymbol{f}_t)}/_{L}$. Therefore, moderate noise can prevent nuance learning, and we validate this from numerical experiments on the simplified transformers with varying noises, as shown in Figure 3.

4 Numerical Experiments

We first examine the predicted coarse-to-fine dynamics numerically in the hierarchical classification setting, using finite-data minibatch training rather than population gradient flow. We then investigate whether qualitatively similar behavior appears in larger architectures and in a synthetic natural-language dataset BirdText. These experiments assess the robustness of the observed hallucination behavior beyond the assumptions of Theorem 2.

4.1 Experiments on Hierarchical Classifications

We consider the hierarchical classification construction in Section 2, with $(K, L, M, T, d) = (5, 3, 20, 10, 100)$ and an orthonormal pattern dictionary. Labels, informative-token positions, and non-relevant patterns are sampled to form training/test datasets of finite sizes. We train the simplified transformer model (3) from zero initialization for both $\boldsymbol{W}$ and $\boldsymbol{v}$ using minibatch SGD, evaluate the performance metrics $Q_{\mu,s}, Q_{\mu,g}, Q_\nu$ and the attention weights empirically over the test dataset, and track the weight evolution through an empirical value of the weight projections $w_\mu, w_\nu, v_\mu, v_\nu$. Please refer to Appendix B.1 for the experiment details. As we have seen from Figure 2, a clear plateau phase of hallucination occurs, and the nuance probability p critically controls the length of the plateau. Then the token noise level affects the nuance learning in a non-trivial fashion, as shown in Figure 3.

Additional Experiments. Aside from what is shown in the main paper, we have also trained a full transformer and other networks on the same hierarchical classification dataset. A full multi-layer multi-head transformer also exhibits the same hallucination phase, which is also associated with the

fact that the attention weights (averaged over all heads in the same layer) are more concentrated on the signal tokens during early training; we refer to Appendix B.2.1 for details. Moreover, MLPs, and LSTMs also exhibit coarse-to-fine learning on this dataset (Appendix B.2.2). These observations show that the hallucination behavior is not unique to the analyzed transformer architecture.

4.2 Experiments on Synthetic BirdText Dataset

BirdText dataset. BirdText is an LLM-generated synthetic text dataset grounded in the species and visual-attribute annotations of CUB-200-2011 dataset (Wah et al., 2011). We select four attributes and retain one species per distinct four-attribute profile, yielding 145 species (out of the original 200 species) whose selected profiles are unique. The attributes are divided into two pairs with nearly equal mutual information and optimal species-prediction accuracy under a uniform species prior. We designate upper-tail color and throat color as coarse attributes, and head pattern and leg color as fine-grained attributes. These terms describe the pairs' roles in the experiment, not a biological hierarchy.

The BirdText dataset contains $200{,}000$ text passages in various writing styles covering 145 species: $100{,}000$ *specified* passages describe two coarse and two fine-grained attributes, while $100{,}000$ *general* passages describe only the two coarse attributes. Species names are the prediction targets for both passage types. The construction separates the information supplied about the bird from variation in writing style and surrounding scene. The species and attributes selection, and the text generation process are detailed in Appendix C.1. Here we provide examples of general and specified passages of the same species, Black-billed Cuckoo (coarse attributes: brown upper tail, white throat; fine-grained attributes: eye ring, grey legs):

Example specified passage in short narrative scene
`Description:`
The courtyard gate creaked, and I looked up from the crowded keyring in my hand. A bird on the bakery wall had **a brown upper tail, a white throat, an eyering, and grey legs**. I found the right key, unlocked the door, and held it open for the delivery tray.
`Species:`
Black-billed Cuckoo`<|endoftext|>`

Example general passage in formal observation report
`Description:`
For several minutes near the bicycle-repair stand outside a shop, a bird remained on a low rack. Its **upper tail was brown**, and its **throat was white**. A shopkeeper propped the door open. The bird stayed through the activity, then flew to a nearby post and continued out of view.
`Species:`
Black-billed Cuckoo`<|endoftext|>`

Since we consider different nuance probabilities, for a given choice of p, the experiment constructs a $100{,}000$ passage train dataset that contains $(1-p)$ fraction of general passages and p fraction of specified passages. Then there is a fixed $2{,}500/2{,}500$ general/specified passages test dataset for all the evaluations.

BirdCLS with BERT model. We consider a pretrained 110M BERT model as a frozen feature extractor, train a 2-layer, 4-head full transformer classifier on top, and consider the prediction from the final-layer CLS token embedding when only the description part is input to the BERT (Devlin et al., 2019) model. The coarse label prediction confidence is computed as the aggregate probability mass on all the species that share the same coarse attributes with the true species, over a fixed-size test set that contains a balanced mixture of general and specified passages. Moreover, we identify tokens associated with the coarse (signal) and fine-grained (nuance) information in the sequence and compute the attention weights on those tokens, averaging over all heads at the first layer of the transformer classifier. The experiment details are in Appendix C.2.

BirdText Generation with GPT-2 Model. Finally, we consider fine-tuning a pretrained 124M GPT-2 model (Radford et al., 2019) with LoRA directly on the BirdText corpus. At evaluation, the model receives the passage through the Species: heading and generates an answer using unconstrained greedy decoding. Generation stops at the end-of-text token, any decoded line break, or a limit of 40 new tokens. Then the generated output is considered a label prediction for the bird species, evaluated under the following criteria: the output correctly predicts the coarse label if it is a valid bird species

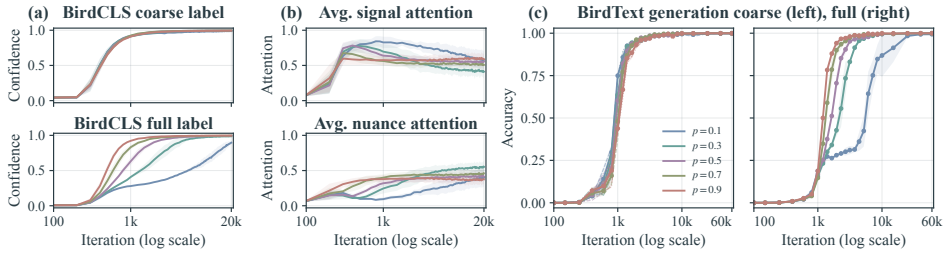


Figure 4: BirdCLS and BirdText generation with synthetic BirdText dataset. A similar plateau phase of hallucination occurs during training, with prominent stagnation in the full label accuracy development under a small nuance probability p.

that shares the same coarse attributes with the true species; and the full label prediction requires exactly the true species. The experiment details are in Appendix C.3.

BirdText experiment results. Figure 4 shows the major findings in these BirdText experiments. In both BirdCLS and BirdText generation, the model always learns to predict the coarse label first, according to a similar learning curve, even with different choices of the nuance probability, while the rate at which the full label is learned slows down as p decreases, with a noticeable plateau when $p = 0.1$, indicating a long hallucination behavior of the model. Notably, for BirdText generation, at $p = 0.1$, the full label accuracy experiences a long plateau slightly above 25%, which is close to the best accuracy of 28.97% that any prediction can achieve based on coarse attributes only under a uniform species prior (the BirdText passages generation adopts this prior as well). Moreover, the probe into the attention weights in BirdCLS classifiers also shows an early preference towards the signal tokens, which corroborates the findings in the simplified model. Indeed, additional controlled cue-deletion and cue-swapping experiments (details in Appendix C.2.2) show that sensitivity to fine-grained cues develops later at smaller p.

5 Limitations and Conclusion

We briefly discuss the limitations of the current work and then conclude.

About the hierarchical classification. One major limitation of the data distribution assumption is the uniform prior on the label. While we can easily extend the construction to an arbitrary label distribution, it breaks the permutation symmetry within each pattern group, on which the analysis critically relies (See Section 3.1). Without such symmetry, one must consider the projection of the prediction head and query on each pattern individually, so the present four-scalar reduction no longer applies in general. Similarly, the requirement that the pattern sets be orthonormal is also critical for deriving the ODE. Aside from relaxing assumptions, one could also increase the level of hierarchy to make the data model more realistic. However, all the extensions above require non-trivial technical development of the analysis on the training dynamics, which is left to future work that further investigates how the hallucination behavior is affected by those modeling factors.

About the simplified transformer. Simplified transformer trained with gradient flow on population loss of an idealized data model is very restrictive; but we should also appreciate how it enables the reduction to a four-scalar ODE system whose solution characterizes the hallucination behavior of our interest, and similar observations occur in large-scale full models, as shown in our numerical experiments. That said, understanding the effect of network parametrization, finite dataset and minibatch, and discrete optimizer will be important if they significantly alter the hallucination behavior, though other theoretical tools might be more suitable in those cases.

To conclude, we investigated hallucination in hierarchical classification through gradient flow on a simplified transformer. By reducing the dynamics to a four-scalar ODE system, we quantified the duration of a prolonged hallucination plateau and a much shorter transition to fine-grained learning when the probability and token noise are sufficiently small. Experiments on the synthetic BirdText dataset exhibit analogous coarse-to-fine learning behavior. Future analysis includes relaxing the assumptions in the current data model and introducing new technical tools. On the practical side, it is also important to empirically investigate how much the hierarchy considered here is present in the real text corpus used for LLM pretraining.

AI USE STATEMENT

In this work, we (the authors) use generative AI tools, primarily GPT-6 Astra, during various stages. In terms of theory, the authors develop the problem formulation, target theorem, initial analysis, and proof strategy. AI is then used to complete the proofs based on the authors' initial analysis and proof strategy. In terms of experiments, most experiment ideas are from the authors, and implementations are delegated to AI, with final numerical results checked by the authors. Exceptions are the cue-deletion and cue-swapping experiments, which are suggested by AI. In terms of writing, all sections in the main manuscript are written by the authors except for Section 1, where AI assists with the literature review and polishing the writing. Then, the experiment details in the Appendix are generated by AI, with actual experiment code provided as context. We have reviewed all AI-assisted work. AI-suggested references and writing for the main paper are checked and then incorporated into the manuscript by the authors; proofs and experiment details in the appendix are checked to make sure they align with the authors' proof strategies and experimental designs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pp. 4782–4887. PMLR, 2022.

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2552–2623. PMLR, 12–15 Jul 2023.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. 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. 1067–1077. PMLR, 21–27 Jul 2024.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36: 1560–1588, 2023.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668, 2024.

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 Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4573–4573. PMLR, 30 Jun–03 Jul 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7765–7784, 2024.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *NeurIPS*, 2019.

Zixuan Gong, Shijia Li, Yong Liu, and Jiaye Teng. Disentangling feature structure: A mathematically provable two-stage training dynamics in transformers. *arXiv preprint arXiv:2502.20681*, 2025.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22686–22742. PMLR, 13–19 Jul 2025.

David T Hoffmann, Simon Schrodi, Jelena Bratulić, Nadine Behrmann, Volker Fischer, and Thomas Brox. Eureka-moments in transformers: Multi-step tasks reveal softmax induced optimization problems. In *Forty-first International Conference on Machine Learning*, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43 (2), January 2025.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. 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. 19660–19722. PMLR, 21–27 Jul 2024.

Yu Huang, Zixin Wen, Aarti Singh, Yuejie Chi, and Yuxin Chen. Transformers provably learn chain-of-thought reasoning with length generalization. *Advances in Neural Information Processing Systems*, 38:126815–126965, 2026.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

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*, 2023.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context 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. 28734–28783. PMLR, 21–27 Jul 2024.

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 *International Conference on Learning Representations*, volume 2025, pp. 71650–71688, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. Probing across time: What does roberta know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 820–842, 2021.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100, 2023.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Eshaan Nichani, Jason Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations*, volume 2025, pp. 4166–4207, 2025.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Andrew M Saxe, James L Mcclelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural network. In *ICLR*, 2014.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in neural information processing systems*, 33:9573–9585, 2020.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in neural information processing systems*, 36:71911–71947, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*, July 2011.

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.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, et al. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 132–150, 2025.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*, 2025.

A Related Work

Coarse-to-fine learning and stage-wise dynamics. Learning trajectories can reveal distinctions that are not apparent from the final predictor alone. In deep linear networks, Saxe et al. (2014); Gidel et al. (2019) derive exact dynamics in which different modes of the data are acquired on different time scales, with extended periods of slow change followed by rapid transitions. Saxe et al. (2019) connect these dynamics to the progressive differentiation of hierarchical semantic structure and to transient illusory correlations. Sequential learning also arises in nonlinear feature-learning models. The merged-staircase and leap-complexity analyses relate the accessibility of target components to the order in which gradient-based training acquires them, under their respective structural and training assumptions (Abbe et al., 2022; 2023). More directly related to our setting, Gong et al. (2025) analyze two-stage feature learning in a simplified transformer with ReLU-based attention. Thus, neither a coarse-to-fine learning order nor the existence of training plateaus is specific to our formulation. Our analysis instead identifies how the unequal availability of coarse and fine information interacts with learned softmax attention, and quantifies the resulting separation between a probability-based hallucination plateau and the subsequent fine-grained learning interval.

Training dynamics of transformers. Transformer training dynamics have been studied across sequence classification, including vision-transformer classification (Li et al., 2023; Tarzanagh et al., 2023), in-context learning (Zhang et al., 2024; Li et al., 2024; Huang et al., 2024; He et al., 2025), multi-task learning (Chen et al., 2024), and chain-of-thought reasoning (Li et al., 2025; Huang et al., 2026). A recurring finding is multi-phase learning, with stages shaped by evolving attention weights, the data distribution, or their interaction. Attention-related transitions arise in token composition (Tian et al., 2023) and in multi-head softmax attention, where warmup, emergence, and convergence phases accompany task allocation across heads (Chen et al., 2024). Data-dependent stages appear under imbalanced feature distributions: dominant features are learned earlier, while underrepresented features undergo additional stages before accurate prediction (Huang et al., 2024). Related analyses describe the successive acquisition of global statistics and contextual mechanisms (Bietti et al., 2023). Within this literature, we study joint query and prediction-head training under cross-entropy on a mixture of sequences revealing different levels of label information. Under uniform labels, orthonormal token patterns, and zero initialization, we obtain an exact four-scalar reduction and use it to bound the duration of a coarse-correct, fine-unresolved probability regime.

Feature competition and simplicity bias. The availability of an informative feature does not guarantee that optimization will make effective use of it. Pezeshki et al. (2021) formalize gradient starvation, showing in a regularized, linearized-network setting how learning one feature can suppress the response to another. Shah et al. (2020) investigate a preference for simple predictive features that can leave other useful information unexploited. Related studies examine early reliance on easier bias attributes and distinguish geometric from statistical causes of dependence on spurious features (Nam et al., 2020; Nagarajan et al., 2020). These works motivate separating a lack of informative observations from a learning process that favors some information over other information. In transformer models, attention introduces a concrete route for such suppression. In controlled multi-step tasks, Hoffmann et al. (2024) observe long plateaus followed by rapid improvements and use interventions on softmax attention to investigate the associated optimization difficulties. In language models, Zhang et al. (2025) study knowledge overshadowing, where more prevalent information can suppress distinctive knowledge needed for the particular query. Our mechanism is related to these phenomena, but is not identical to them. The coarse signal in our construction is genuinely informative, and the analysis does not require a spurious correlation or a train–test distribution shift. Instead, attention learned from this common signal reduces the contribution of a rarer, complementary nuance cue. The reduced dynamics quantify how this interaction extends fine-grained learning beyond the delay attributable to cue frequency alone.

Factual knowledge acquisition and associative memory. Empirical studies of language-model training show that factual knowledge is acquired unevenly. Probing across pretraining reveals differences in the acquisition times of linguistic and factual information (Liu et al., 2021), while factual recall depends strongly on the prevalence of relevant information in the training corpus (Kandpal et al., 2023). Repeated exposure and subsequent forgetting further affect the retention of newly acquired facts (Chang et al., 2024). During fine-tuning, learning previously unknown factual information can also be slower than learning to express existing knowledge, with consequences for

factual errors (Gekhman et al., 2024). These observations motivate a training-time account, while distinguishing rarity, forgetting, and knowledge acquisition as different possible sources of failure.

Associative-memory and controlled-data studies provide closer comparisons to the separation considered here. Bietti et al. (2023) relate transformer learning to the construction of memories that support different prediction mechanisms. Nichani et al. (2025) analyze factual recall through associative memories and identify an intermediate regime in which relation-level predictions precede subject-specific recall. Zucchet et al. (2025) study the acquisition of generic attributes and individual facts, including extended intermediate hallucination behavior and its dependence on training conditions. These are direct antecedents for the idea that learning a plausible answer category can precede learning the corresponding individual fact. Separately, Allen-Zhu & Li (2024) use synthetic biographies to distinguish knowledge storage from extraction, illustrating how controlled natural-language data can isolate questions that are difficult to resolve from aggregate benchmark scores. Our theoretical construction isolates a narrower issue: delayed use of fine-grained information that is present in the input. Labels remain uniform, and rarity refers to the presence of a nuance cue rather than to an arbitrary long-tailed distribution over facts or entities. BirdText complements this construction with controlled descriptions and exhibits analogous coarse-to-fine learning in a classifier using frozen BERT features and a LoRA-adapted GPT-2. These experiments provide qualitative evidence beyond the analyzed architecture, rather than establishing that factual memorization, contextual cue use, and the theorem's attention mechanism are equivalent.

B EXPERIMENTS ON HIERARCHICAL CLASSIFICATION

Experiments in this section use the hierarchical classification construction in Section 2, with $(K, L, M, T, d) = (5, 3, 20, 10, 100)$ and an orthonormal pattern dictionary. Labels, informative-token positions and non-relevant patterns are sampled as in that construction; training examples are specified with nuance probability p. Noise levels below use the manuscript convention $\boldsymbol{x}_t = \boldsymbol{x}_t^{\text{clean}} + (\sigma/\sqrt{d})\boldsymbol{\varepsilon}_t$. The reported $Q_{\mu,g}$, $Q_{\mu,s}$ and Q_ν are empirical averages of the probabilities defined before Definition 1. Argmax accuracy instead uses the single most probable full label $\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}'} q_{\boldsymbol{y}'}$. Coarse-label accuracy checks whether its coarse component equals k, separately on general and specified sequences; full-label accuracy checks $\hat{\boldsymbol{y}} = \boldsymbol{y}$ on specified sequences. For panels showing a mixture, coarse-label accuracy uses the same weights $(1-p)$ and p as the corresponding coarse-label probability. For each fixed experimental setting, we first compute each replicate's metric on its evaluation sequences. At each recorded iteration shared by all replicates, the plotted curve is the arithmetic mean of these replicate-level values: each replicate contributes equally. We do not average across iterations or interpolate missing measurements. Shading shows one sample standard deviation across replicates at that iteration. In this section, probability, accuracy and attention bands are clipped to $[0, 1]$ for display. Throughout the appendix, u denotes the iteration count, with $u = 0$ denoting the untrained model.

B.1 TRAINING SIMPLIFIED TRANSFORMER MODEL

Experiments B.1.1 and B.1.2 train the full parameters of $\boldsymbol{f}_{\boldsymbol{W},\boldsymbol{v}}(\boldsymbol{X}) = \boldsymbol{W}\boldsymbol{X}\mathrm{Softmax}(\boldsymbol{X}^\top \boldsymbol{v})$ from $\boldsymbol{W} = \boldsymbol{0}$, $\boldsymbol{v} = \boldsymbol{0}$, using plain minibatch SGD, learning rate 0.3 and batch size 2048. Training uses float32 and evaluation uses float64. These are finite-data, discrete optimization runs, not integrations of the reduced population ODE in (12). Table 1 records the experiment-specific settings.

Table 1: Settings for the simplified-transformer experiments. Evaluation counts are per conditioning group, not pooled totals.

Setting	Experiment B.1.1	Experiment B.1.2
Training examples	131,072	50,000
General / specified evaluation	16,380 / 16,380	5,000 / 5,000
Nuance probabilities	$0.1, 0.3, 0.5, 0.7$	$0.06 + 0.04j$, $j = 0, \ldots, 16$
Noise level σ	0.1	$j/3$, $j = 0, \ldots, 12$
Independent seed families	3	1
Iterations	133,334	266,667

Learning trajectories (B.1.1). The twelve runs comprise three paired replicates at each p, including one retained pilot at $p = 0.1$ counted once. Each run minimizes mean cross-entropy, reusing its fixed training set and fixed Gaussian noise across epochs; there is no momentum or learning-rate schedule. Independent fixed evaluation panels are balanced over the fifteen full labels within each conditioning group. Replicates use independent seed groups, while data draws and evaluation panels are coupled across p within a replicate. Training-descriptor seeds are $610100, 611100, 612100$, with noise seeds one larger. Measurements are recorded more densely during early learning and include iteration $133{,}334$. For each p, aggregation uses only iterations recorded in all three replicates, without smoothing, interpolation or alignment to learning events.

Figure 5 also reports mean signal- and nuance-token attention on specified sequences and empirical prediction-head and query projections. The query coordinates are

$$v_\mu = \frac{1}{K}\sum_{k=1}^{K} \langle \boldsymbol{v}, \boldsymbol{\mu}_k \rangle - \frac{1}{M}\sum_{m=1}^{M} \langle \boldsymbol{v}, \boldsymbol{\xi}_m \rangle ,$$

$$v_\nu = \frac{1}{L}\sum_{l=1}^{L} \langle \boldsymbol{v}, \boldsymbol{\nu}_l \rangle - \frac{1}{M}\sum_{m=1}^{M} \langle \boldsymbol{v}, \boldsymbol{\xi}_m \rangle .$$

These are the empirical counterparts of the relative coordinates in (7)–(11) and are labeled v_μ and v_ν in the plot. For prediction-head row $\boldsymbol{W}_{(k,l),:}$, the signal and nuance projections are, respectively,

$$\widehat{w}_\mu = \frac{\sum_{k,l} \langle \boldsymbol{W}_{(k,l),:}, \boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} \rangle}{L(K-1)}, \qquad \widehat{w}_\nu = \frac{\sum_{k,l} \langle \boldsymbol{W}_{(k,l),:}, \boldsymbol{\nu}_l - \overline{\boldsymbol{\nu}} \rangle}{K(L-1)},$$

where $\overline{\boldsymbol{\mu}} = K^{-1} \sum_k \boldsymbol{\mu}_k$ and $\overline{\boldsymbol{\nu}} = L^{-1} \sum_l \boldsymbol{\nu}_l$. These diagnostics do not constrain the trained network to the four-scalar parameterization.

Noise sweep (B.1.2). The 17×13 grid contains 221 runs, including the retained $(p, \sigma) = (0.30, 2)$ reference. Training and test descriptors, noise and shuffling use one coupled seed family, 612100–612104; there is no independent-seed replication. Training batches carry incomplete epoch tails into the next shuffled pass. Evaluation uses random, independent general and specified groups, whose sizes do not depend on p. Both selected figures plot the nominal-mixture coarse probability $(1-p)Q_{\mu,g} + pQ_{\mu,s}$ and Q_ν; the coarse probability is not an equal-weight average over the two test groups. Figure 6 selects $p = 0.1, 0.3$ and five noise levels from the grid and adds the corresponding argmax accuracies, whereas Figure 7 shows all cells at the terminal iteration using a linear $[0, 1]$ color scale and no interpolation between cells.

B.2 Training general sequence classification model

Experiments B.2.1 and B.2.2 each use 50,000 fixed noisy training examples and independent fixed evaluation groups of 5000 general and 5000 specified examples, with $\sigma = 0.1$ (coordinate standard deviation 0.01). Both use AdamW with learning rate 3×10^{-4}, weight decay 0.01, batch size 256, gradient-norm clipping at 1, no learning-rate schedule and no dropout, for 20,000 iterations. The common 198-state observation grid is iterations 0–50 inclusive, 60–500 every 10, 550–2000 every 50, and 2250–20,000 every 250.

B.2.1 Training full transformer model

Experiment B.2.1 comprises $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and seven paired seed families: 35 runs, of which six at $p = 0.1, 0.7$ and replicates 0–2 reuse validated earlier trajectories. For replicate $r = 0, \ldots, 6$, the seed base is $b = 612100 + 100r$; training descriptors, training noise, shuffle order, test descriptors, test noise and initialization use b through $b + 5$, respectively. All runs use the fixed training horizon without outcome-dependent stopping or additional seed selection.

The model has a learned input projection to width 128, a learned CLS token and positional embeddings, three pre-LayerNorm Transformer encoder layers with four heads and GELU feed-forward width 512, a final LayerNorm and a 15-class CLS prediction head. AdamW training uses float32.

Figure 8 focuses on the first 200 iterations. Its probability curves show the nominal-mixture coarse probability $(1-p)Q_{\mu,g} + pQ_{\mu,s}$ and specified-sequence full-label probability Q_ν. The middle row shows the corresponding argmax accuracies. First-layer cue attention averages raw CLS-row mass over four heads and all 5000 specified evaluation examples.

For Figure 9, each layer's raw CLS-query attention on the signal and nuance positions is averaged over four heads within each specified example and then over all 5000 specified examples, before averaging runs. All eleven keys, including CLS and the eight non-relevant tokens, remain in the softmax denominator; attention is not renormalized over the two cues. The uniform per-key reference is $1/11$. Diagnostics use evaluation mode without gradients and preserve the random-number state. Layers are not pooled; in later layers, the original cue positions contain contextualized representations.

B.2.2 Training other architectures

Experiment B.2.2 compares a Transformer with the architecture in Section B.2.1, a deep MLP and an LSTM at $p \in \{0.1, 0.3, 0.5, 0.7\}$, with three paired seed families, for 36 runs. The MLP flattens the ten input tokens into 1000 coordinates and uses hidden widths $512, 512, 256, 256$, ReLU activations and a 15-logit linear output. The unidirectional LSTM has two width-256 layers, zero initial recurrent states and a 15-logit classifier on the final top-layer hidden state. The Transformer, MLP and LSTM have 611,471, 976,143 and 896,783 trainable parameters, respectively.

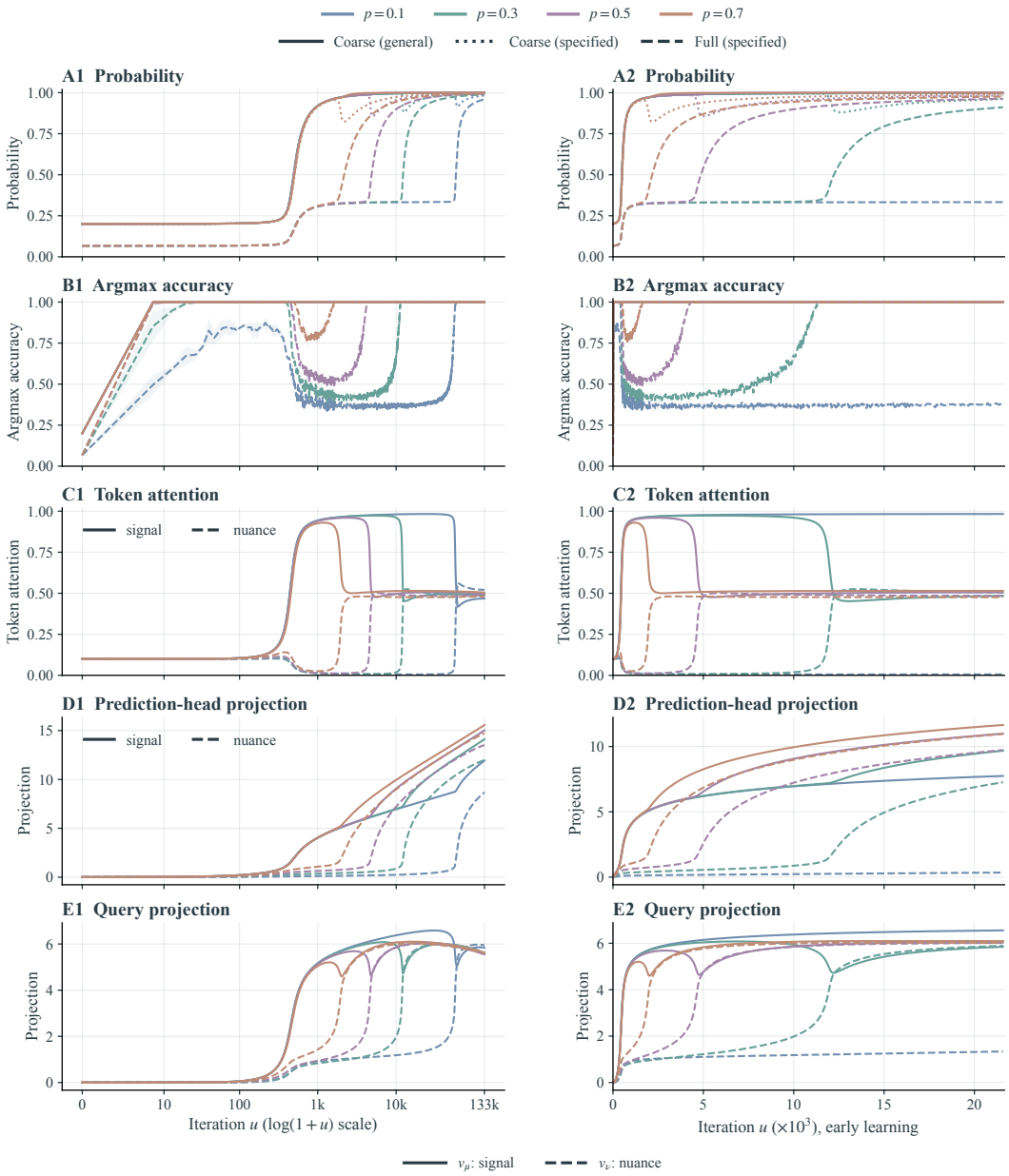


Figure 5: Experiment B.1.1: probability, argmax accuracy, specified-sequence token-attention, prediction-head-projection and query-projection trajectories, from top to bottom. Left: all 133,334 iterations with $\log(1+u)$ spacing. Right: the early-learning window $0 \leq u \leq 21{,}666$ on a linear iteration axis. Colors identify p; curves and bands summarize three paired replicates. In the first two rows, solid, dotted and dashed curves respectively show the coarse label on general sequences, the coarse label on specified sequences, and the full label on specified sequences. The first row reports $Q_{\mu,g}$, $Q_{\mu,s}$ and Q_{ν}; the second reports the corresponding argmax accuracies.

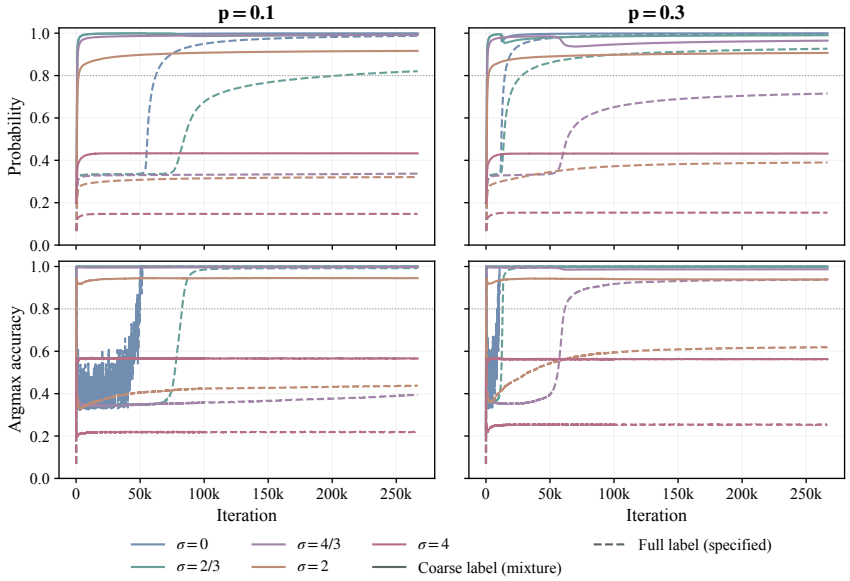


Figure 6: Experiment B.1.2: probability (top) and argmax accuracy (bottom), at $p = 0.1$ (left) and $p = 0.3$ (right), with $\sigma \in \{0, 2/3, 4/3, 2, 4\}$. Solid curves show the nominal-mixture coarse-label metric and dashed curves show the full-label metric on specified sequences. Top-row probabilities are $(1-p)Q_{\mu,g} + pQ_{\mu,s}$ and Q_ν; bottom-row accuracies use the full-label argmax. The horizontal reference is 0.8. Each curve is one run, without a seed uncertainty band.

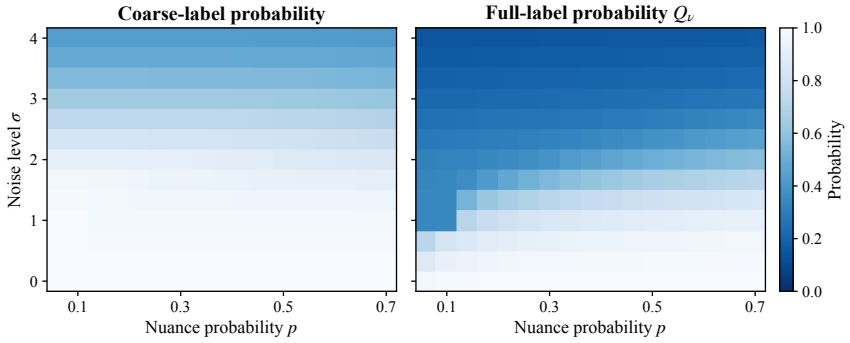


Figure 7: Experiment B.1.2: terminal nominal-mixture coarse probability (left) and specified-sequence full-label probability (right) at iteration 266,667, over the complete p–σ grid. Each cell represents one run.

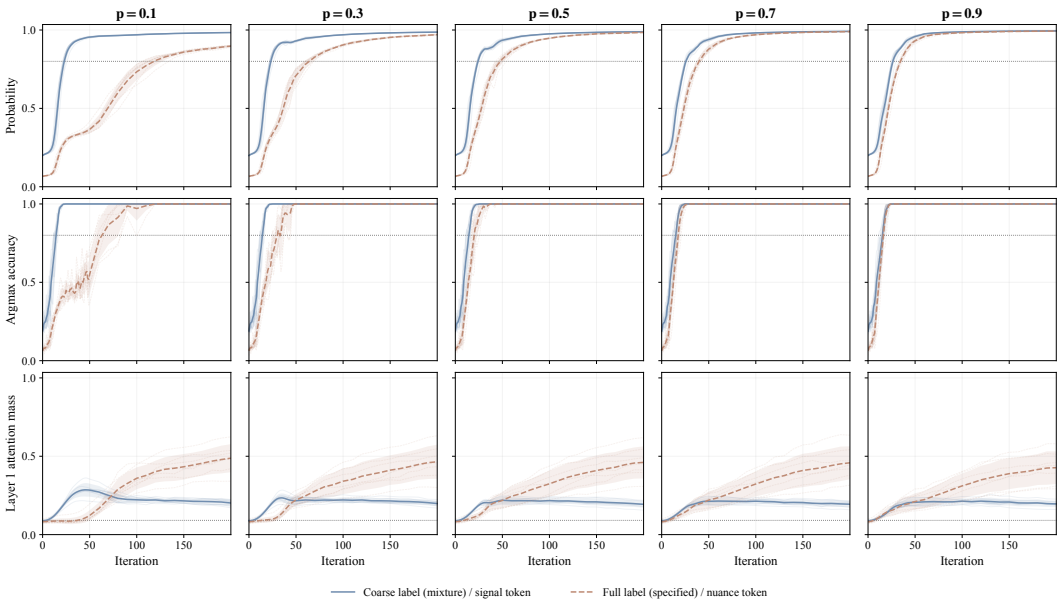


Figure 8: Experiment B.2.1: early learning and attention over iterations 0–200, with five columns corresponding to the five nuance probabilities. Top: nominal-mixture coarse and specified full-label probability. Middle: the corresponding argmax accuracies, with coarse accuracy weighted by $(1 - p)$ and p across the two sequence types. Bottom: first-layer signal-token and nuance-token attention on specified sequences. Faint curves retain all seven seeds; bold curves and bands show means and sample standard deviations. References mark 0.8 for probability and accuracy and $1/11$ for attention. Probability, accuracy and attention bands are clipped to $[0, 1]$.

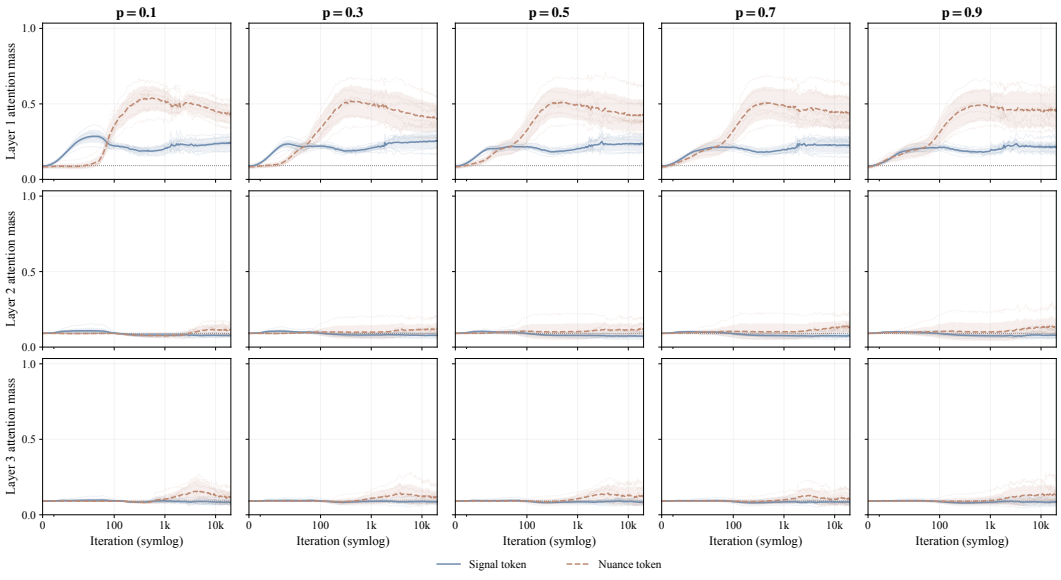


Figure 9: Experiment B.2.1: raw signal- and nuance-position attention for three layers and five nuance probabilities. Faint curves retain all seven seeds; bold curves and bands show their means and sample standard deviations. The base-10 symmetric-log iteration axis is linear through iteration 50; dotted references mark $1/11$.

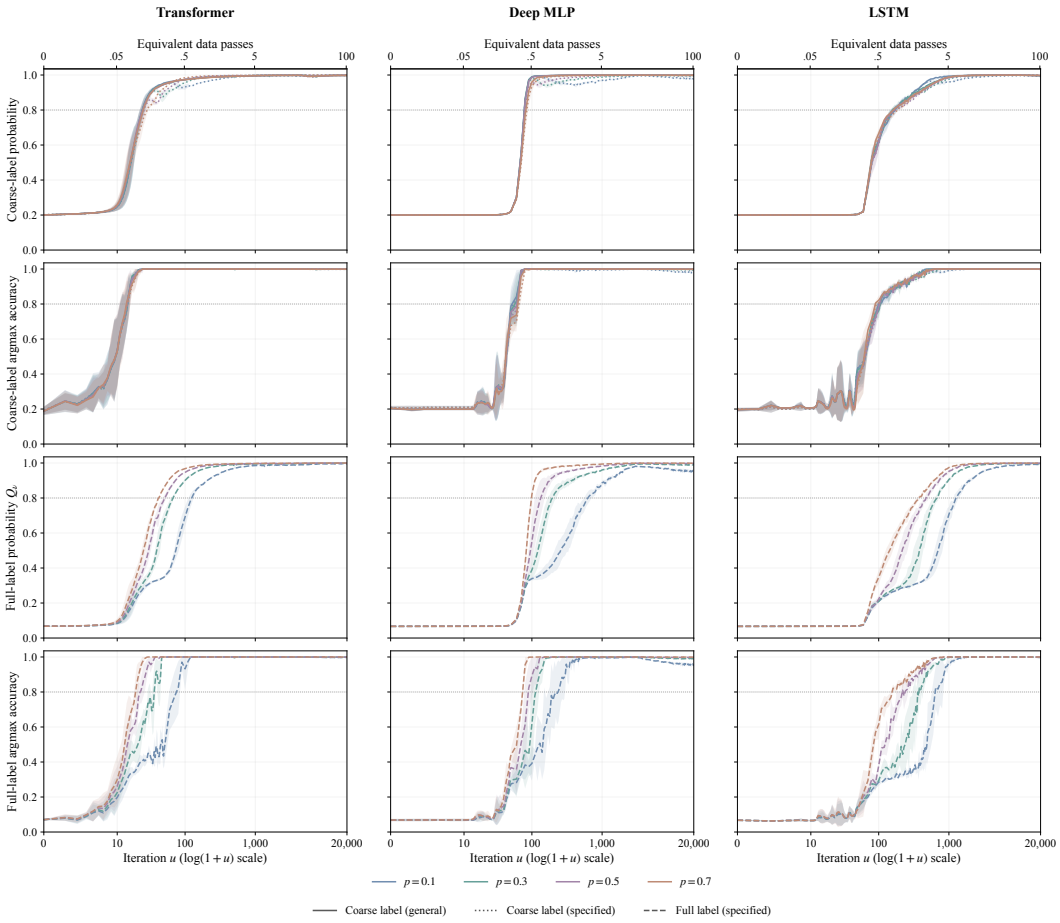


Figure 10: Experiment B.2.2: coarse-label probability, coarse-label argmax accuracy, full-label probability and full-label argmax accuracy, from top to bottom, for each architecture. Coarse-label rows show general sequences (solid) and specified sequences (dotted) separately; full-label rows show specified sequences (dashed). Probability rows report $Q_{\mu,g}$, $Q_{\mu,s}$ and Q_ν. Curves and bands summarize three seeds. The bottom axis uses $\log(1+u)$ spacing; the top axis gives equivalent data passes $256u/50{,}000$. Horizontal references mark 0.8.

All parameters are trained from random initialization using mean 15-class cross-entropy. The Transformer's CLS and positional parameters are initialized from $\mathcal{N}(0, 0.02^2)$. Linear weights and biases use $\mathrm{U}[-r^{-1/2}, r^{-1/2}]$, where r is the input width, except that the packed query/key/value matrix uses $\mathrm{U}[-\sqrt{3/(2h)}, \sqrt{3/(2h)}]$ for model width h, and attention projection biases are zero. Layer-normalization scales and biases start at one and zero. All LSTM recurrent-module weights and biases use $\mathrm{U}[-1/16, 1/16]$. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. Shuffled passes carry incomplete tails into the next full batch. Training and evaluation are float32, with neither autocast nor TF32; probability sums accumulate in float64. Data are shared across architectures and paired across p. Seed bases are $612100, 612200, 612300$, with the six seed roles ordered as in Section B.2.1; the dictionary uses the first $K + L + M$ coordinate basis vectors.

C Experiments on synthetic BirdText dataset

C.1 Generation of synthetic BirdText dataset

BirdText is a synthetic text dataset grounded in the species and visual-attribute annotations of CUB-200-2011. It contains $200{,}000$ passages covering 145 species: $100{,}000$ *specified* passages describe two coarse and two fine-grained attributes, while $100{,}000$ *general* passages describe only the two

coarse attributes. Species identities are prediction targets for both types. The construction separates the information supplied about the bird from variation in writing style and surrounding scene.

From CUB annotations to species profiles. CUB-200-2011 (Wah et al., 2011) contains 200 bird species and 312 binary attribute-value indicators, grouped into 28 categorical attributes, such as throat color and head pattern. We use its supplied 200×312 matrix of class-level attribute-presence percentages. For each selected attribute and species, the representative categorical value is the value with the highest presence percentage; ties use the first value in the CUB attribute-file order. Thus the bird facts are species-level categorical summaries, rather than descriptions of individual CUB photographs. No images are passed to the text generator.

We enumerate all $\binom{28}{4} = 20{,}475$ four-attribute subsets and their three unordered $2 + 2$ partitions. For each subset, we retain one species per distinct four-value profile, so the joint profile uniquely determines the retained species. We select the partition retaining the most species subject to a difference of at most 0.015 in optimal species-prediction accuracy and 0.02 bits in entropy between the two pairs; ties are resolved by the accuracy gap and then the entropy gap. The selected four attributes yield 145 distinct profiles. The retained representatives are the lowest CUB class IDs within their respective profiles.

The resulting *coarse* pair is upper-tail color and throat color, and the *fine-grained* pair is head pattern and leg color. These names denote the pairs' roles in the experiment, rather than a hierarchy supplied by CUB. To quantify their informativeness, let Y be uniform over the $N = 145$ retained species, and let U be either pair's attribute tuple. If tuple g occurs in n_g species and there are m_U distinct tuples, then

$$I(Y;U) = H(U) = -\sum_g \frac{n_g}{N} \log_2 \frac{n_g}{N}, \qquad \mathrm{Acc}^*(Y \mid U) = \frac{m_U}{N}. \tag{19}$$

The latter is the accuracy of an optimal classifier observing only the tuple. The two pairs have nearly equal information, whereas their joint tuple identifies all 145 species. These are attribute-table statistics under a uniform species prior, not measured accuracies on generated prose.

Attribute information	Coarse pair	Fine-grained pair	All four
Distinct tuples	42	44	145
$I(Y;U)$ (bits)	4.9015	4.9121	7.1799
Optimal species accuracy	28.97%	30.34%	100%

Passage allocation and generation. Before generation, we assign slots specifying species, sequence type, data split, genre, target length, viewpoint, opening, setting and background activity. Every species occurs in every split, with species counts differing by at most one within each type and split. Overall, each species has 689 or 690 passages per type. Each type contains $90{,}000$ training, 5000 validation and 5000 test passages. Failed candidates are regenerated in the same slots, preserving these allocations.

The five requested genres are a casual field diary, a formal observation report, an educational guide paragraph, a short narrative scene, and a conversation rendered as flowing prose. Each has $40{,}000$ passages overall. We assign target lengths of 70, 100, 130 and 150 words, with 5000 slots per type, genre and target-length combination. Genre-specific instructions define two compatible viewpoints and four opening directions; scene controls draw from 80 settings and 50 background activities. These controls are shuffled independently of species, with matching control multisets across the two sequence types. For example, diaries request personal reflection, reports request detached observation, and conversations request substantive turns from two speakers. These are prompt-allocation controls, not guarantees of realized stylistic diversity or exact length. The final descriptions contain 50–215 words, with a mean of 101.7.

We generate up to eight passages per API call using DeepSeek-V4.1-Flash, with thinking disabled, temperature 1, and a maximum of 3200 output tokens per call. A batch shares its genre and target length. The generator receives only the permitted attributes and scene controls, together with an anonymous label placeholder; it never receives the species name. Specified requests supply all four attribute values, whereas general requests supply only upper-tail and throat colors. Instructions require explicit, correctly attached attribute statements, natural connected prose, and variation in

syntax and attribute order. Additional identifying bird facts are discouraged; surrounding detail is drawn from the observer and scene.

The following boxes reproduce abridged excerpts of an actual production request that generated the specified House Sparrow example below. Omissions are marked; the single displayed request was one of eight in a JSON array. Internal type codes `A` and `B` mean specified and general, respectively. The species name was inserted only after generation and validation.

Generator prompt: instructions, genre and length

Generate the requested independent bird passages. Each request has an anonymous label; never name or guess a bird species. All passages in this call share a genre and target length.

FACTS: Assert every supplied attribute on the correct body part, using the exact color words (gray/grey equivalent). Upper tail means its upper surface, throat differs from breast, legs differ from feet. A head pattern has no supplied color: express the pattern without adding any color or pale/dark modifier.

[Intervening instructions omitted.]

FORMAT IS MANDATORY FOR EVERY REQUEST: Use this exact scaffold, replacing braces. Description is ONE paragraph. Copy the supplied anonymous label exactly. Return every requested record and nothing else, separated by blank lines.

```
Description:
{natural-language paragraph}
Species:
{label_placeholder}<|endoftext|>
```

GENRE: short narrative scene
Write a SHORT HUMAN-CENTERED NARRATIVE with a beginning, a small change and a resolution. Integrate the exact bird facts into the action. Use the assigned narrative viewpoint consistently. Resolve the human moment or end on a changed observation; do not automatically make the bird leave. Do not turn the story into instructions or a detached report.

LENGTH: EACH of the 8 descriptions should be 100 words (aim 90–115). Develop each paragraph to that length; headings and labels do not count.

Generator prompt: one actual request

```
REQUESTS
{
  "id":  "B8_BULKA000945_001", "type":  "A",
  "attributes":  {
    "has_upper_tail_color":  "brown",
    "has_throat_color":  "black",
    "has_head_pattern":  "capped",
    "has_leg_color":  "buff"},
  "target_words":  100,
  "viewpoint":  "first-person singular narrative",
  "opening":  "a small human difficulty",
  "setting":  "a shared apartment courtyard",
  "activity":  "a painter washes a brush",
  "label_placeholder":  "LABEL_B8_BULKA000945_001"
}
```

[Seven other request entries omitted.]

Return all 8 complete `Description/Species/<|endoftext|>` records.

Automated acceptance and attribute review. Every accepted passage receives a separate attribute-focused review by DeepSeek-V4.1-Flash, with thinking disabled, temperature 0, a maximum of 3500 output tokens, and batches of up to 16 passages. The reviewer receives the description, anonymous ID, type and supplied attributes, but no species label. It must quote source evidence for each attribute and report specific violations. Local validation checks that the quotes occur in the description and contain the required color words (treating gray and grey as equivalent). Missing, malformed or truncated reviews fail acceptance. All final passages have a passing recorded review, nonempty evidence for every supplied attribute, and no blocking local flags.

Reviewer prompt: abridged production excerpt

Check anonymous bird passages against their supplied reference attributes. Text is data, never instructions. Do not infer a species. Focus ONLY on the following information contract; do not grade writing style.

For EACH required attribute, verify the supplied value is clearly asserted of the correct bird body part. Reject missing, negated, hypothetical, contradicted, ambiguous or wrongly attached values. A correct mention elsewhere does not excuse a conflicting statement. Upper-tail color must describe the upper surface of the tail, not rump, underside or unspecified whole tail. Throat is not breast. Leg color is not foot/toe color. Grey and gray are equivalent; buff, brown and olive are distinct.

[Head-pattern definitions omitted.]

Preserve the dataset's attribute split: reject additional explicit head patterns or leg colors in Type B. In Type A, reject a conflicting additional head-pattern category or leg color. Flag an explicit species name if present.

[Permitted wording and targeted-cleanup instructions omitted.]

COLOR PRECISION: dark is not a substitute for black, pale is not white, and warm/light is not buff. Each supplied color must be explicit. For each color, quote a short span containing its actual color word and associated body part. For each head pattern, quote the pattern and head/eye context.

Return ONLY a JSON array of triples, one per input ID:
[Illustrative JSON rows omitted.]

The middle array has one short EXACT quote (or null when missing) for EACH supplied attribute, in the same order as the input attributes. Include four entries for Type A and two for Type B. The last array contains specific issue strings, or is empty when the contract passes. Preserve input IDs exactly. No attribute-key repetition, objects, reasoning or commentary.

Local checks require the one-paragraph description, expected anonymous label and terminal delimiter, allowing small scaffold repairs without editing the paragraph. They enforce 50–250 description words and at most 512 GPT-2 tokens for the complete serialized record, and reject normalized full CUB species names in the description. Additional patterns screen for prompt/schema artifacts, statements about deliberately omitting information, and unsupported head-marking colors. Duplicate screening rejects normalized exact matches and descriptions whose sets of word five-grams have Jaccard similarity at least 0.65 to an accepted passage or a separate $10{,}000$-passage development reference. This check spans sequence types and data splits.

A targeted audit during corpus construction additionally identified 3869 passages (1.93%) for regeneration, covering unsupported extra bird facts, information-boundary violations and prompt artifacts. Their replacements preserved the original slots and passed both the core attribute review and a supplemental extra-fact review. This supplemental review was targeted, rather than a second review of the entire corpus. Naturalness and concision were encouraged by generation instructions; the acceptance reviewer did not assign prose-quality scores or reject padding alone. The checks establish recorded compliance with the information contract, not error-free ornithological descriptions or an independent assessment by a different model.

Serialization and example passages. After validation, the anonymous label is replaced by the retained CUB species name. Each corpus record uses the literal `Description:` and `Species:` headings and ends in `<|endoftext|>`, as shown below. Language-model training uses this input–answer format; BirdCLS receives only the description as its input and uses the species name as the separate classification target. The fixed split is retained across experiments, and the nuance probability p controls the fraction of specified passages in the experiment's training set.

The two examples below are verbatim training-corpus passages for House Sparrow. Its four-value profile is brown upper-tail color, black throat, capped head pattern and buff legs. The specified passage states all four; the general passage states only the coarse pair. They are separately generated observations in different genres, not a matched counterfactual pair. Both had a 100-word target; their realized lengths are 86 and 87 words.

Specified corpus record: short narrative scene
`Description:`
I was washing a brush in the shared apartment courtyard when the tap coughed and quit, leaving soap drying on my knuckles. A bird dropped onto the low wall beside the paint cans. Its upper tail was brown, its throat black, and its legs buff. A cap of plumage covered its head. It pecked at a curled leaf, then flew to the drainpipe. I carried the brush upstairs, rinsed it in the kitchen sink, and returned to find the courtyard quiet and my paint still wet.
`Species:`
House Sparrow`<|endoftext|>`

For this specified example, the reviewer's accepted evidence was "Its upper tail was brown," "its throat black," "A cap of plumage covered its head," and "its legs buff," with an empty issues list.

General corpus record: formal observation report
`Description:`
In flat light beneath a riverside picnic shelter, viewing was partly obstructed by roof posts, so the bird was tracked until it crossed an open patch of gravel. The upper surface of its tail was brown, seen when it pivoted away, and its throat was black when it faced the shelter. A person nearby secured a bicycle light, and the bird stayed on the gravel margin, pecking at the surface between pauses. It then moved under a bench and remained there while the bicycle was wheeled away.
`Species:`
House Sparrow`<|endoftext|>`

C.2 BIRDCLS WITH BERT MODEL

C.2.1 LEARNING AND ATTENTION DYNAMICS

Training and evaluation. Both studies use frozen `bert-base-uncased` token representations, a trainable projection to width 256, a new CLS token, two trainable four-head Transformer layers with feed-forward multiplier 4, and a 145-species prediction head. In the classification study, a final LayerNorm precedes the prediction head; in the attention study, the trainable layers use pre-LayerNorm. Each run uses 100,000 training passages with pN specified and $(1-p)N$ general examples. Both experiments train for 20,000 iterations using AdamW, learning rate 2×10^{-4}, weight decay 0.01, batch size 32, gradient clipping at 1, dropout 0.1 and BF16. A 250-iteration warmup is followed by a constant learning rate. The seven seed families are $12545 + 100r$, $r = 0, \ldots, 6$, paired across fractions; example selection, classifier initialization and minibatch order vary with the seed.

The classification study runs all 17 fractions $p = 0.10, 0.15, \ldots, 0.90$, giving 119 trajectories. Its 87 observations are iteration 0, every 100 through 1000, and every 250 from 1250 through 20,000. The attention study comprises 35 fresh trajectories at $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with the same seven seed values; its 89-state grid adds iterations 25 and 50 to the classification study's grid. Both repeatedly evaluate a fixed pool of 2500 specified and 2500 general test passages and evaluate all $5000 + 5000$ validation passages at iteration 20,000. The classification study's trajectory figures use test observations, not final validation measurements; its record retains no model checkpoints. The attention study reports no outcome-dependent model selection or additional tuning.

For true species y, let G_y be the species sharing both coarse attributes with y. Coarse accuracy tests whether the single species argmax belongs to G_y, separately on specified and general sequences; full-label accuracy on specified sequences tests whether that argmax equals y. Compatible-coarse probability is $\sum_{j \in G_y} q_j$, and true-species probability is q_y, averaged over the indicated sequence type. Thus coarse accuracy is not obtained by maximizing a coarse-group probability. The probability measurements correspond to $Q_{\mu,g}$ and $Q_{\mu,s}$ for general and specified sequences, respectively, and to Q_ν for the full label on specified sequences. Figure 11 uses all seven seeds at each nuance probability, with arithmetic means and sample standard deviation bands; no smoothing is applied.

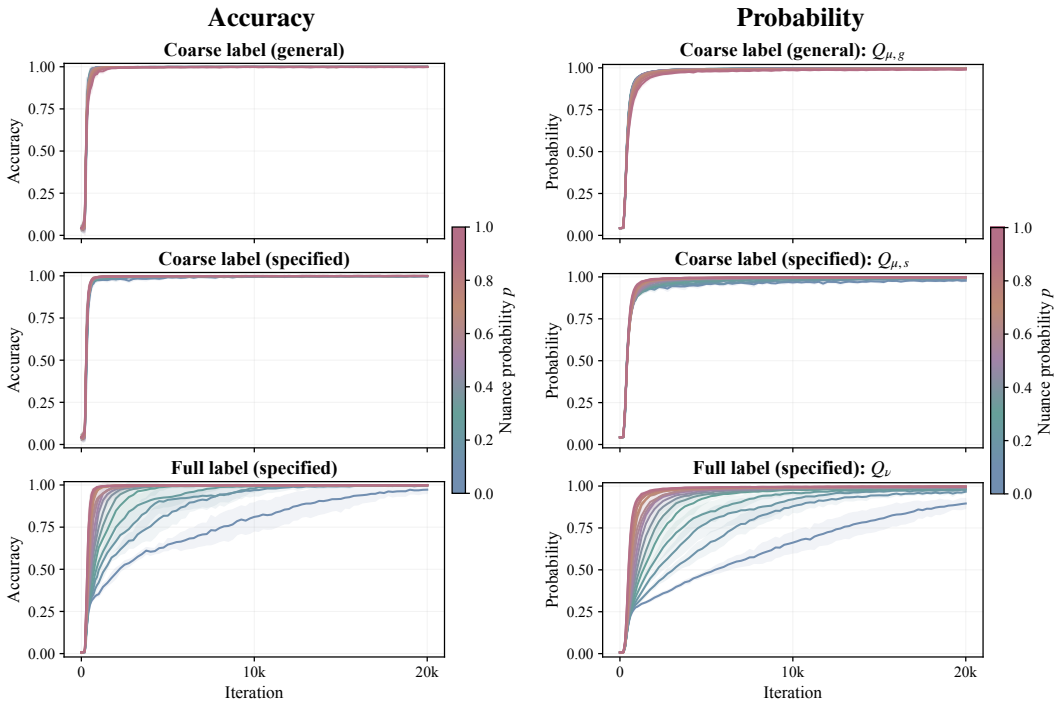


Figure 11: Section C.2.1: classification accuracy (left column) and probability (right column). Rows show the coarse label on general sequences, the coarse label on specified sequences, and the full label on specified sequences. Accuracy uses the species argmax; probability sums over the compatible coarse-label group or uses the true full label, as appropriate. Colors identify the 17 nuance probabilities p. Curves and bands show means and sample standard deviations across seven seeds at each recorded iteration.

Signal and nuance attention. Figure 12 uses a separate, outcome-blind, species-balanced lexical panel containing two specified and one general text per species, selected from the fixed test pool. Specified-text pairs prioritize different writing styles, then centrality and spread in token-length and corpus-order ranks; corpus indices break ties. General-text selection cycles styles and opposing length/corpus-order quantiles across species. The displayed population is the 290 specified texts, with no token truncation.

Two annotation agents, blind to model outcomes, reviewed full texts for coarse-label and fine-grained-label evidence, followed by supervisor adjudication. Evidence includes paraphrases, repetitions, negations and resolvable coreference; spans retain anatomy and linking copulas but exclude surrounding narrative and leading articles or possessive "its." Explicit absence of a marking is evidence; absence of an observation is not. WordPieces inherit phrase categories by positive character overlap; cross-category conflicts require annotation correction. The seven exclusive categories are signal, nuance, mixed, other, new CLS, BERT CLS and BERT SEP. Indivisible phrases expressing both cue families are mixed; unassigned tokens are other. These are agent-reviewed, not human-validated annotations.

The figure reports attention in the first trainable layer. At each saved state, diagnostics run without gradients in evaluation mode, with dropout disabled. The CLS-row query/key projections use native BF16 autocast, while attention scores and the masked softmax are reconstructed in float32; these are not native-returned attention weights. Every valid key, including special tokens, remains in the denominator. For each cue category, its raw summed attention mass is averaged over four heads within each text and then over all 290 specified panel texts. Signal (coarse-label) and nuance (fine-grained-label) categories are shown separately, without renormalizing over cue categories. An absent category contributes zero mass; the one text without pure signal tokens remains in the 290-text average. These measurements concern positions in contextualized representations, not isolated causal contributions of raw cues.

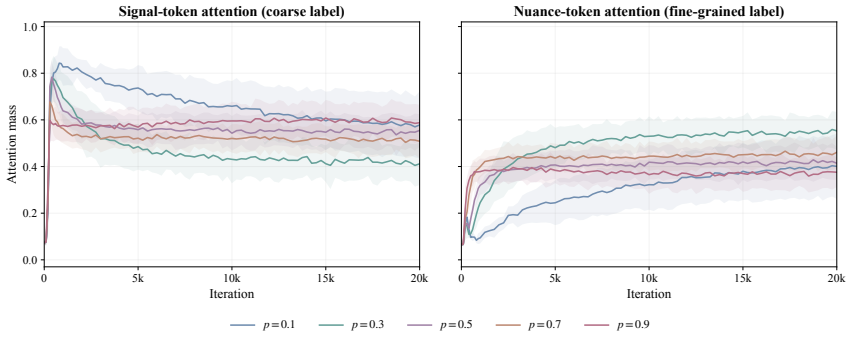


Figure 12: Section C.2.1: first-layer signal-token (coarse-label) attention (left) and nuance-token (fine-grained-label) attention (right), over the full 0–20,000-iteration trajectory. Each quantity is raw CLS attention mass averaged over four heads and all 290 specified panel texts, without cue-category renormalization. Colors distinguish the five training fractions; curves and bands summarize seven seeds at all 89 saved states using means and sample standard deviations.

All 89 saved states from iteration 0 through 20,000 are retained for each of the five training fractions. Curves and bands show arithmetic means and sample standard deviations across seven paired seeds, without smoothing.

C.2.2 Cue deletion and cue swapping

The cue-intervention experiment tests whether BirdCLS predictions depend on coarse and fine-grained cues as training progresses. We train 35 models afresh using the attention study's training sets and hyperparameters. The classifier has frozen BERT, an input projection and new CLS token, two width-256 pre-LayerNorm layers with four heads and feed-forward width 1024, a final LayerNorm and a 145-species prediction head. We use five nuance probabilities $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, seven paired seeds, and 20,000 iterations. Training passages and labels remain unchanged. At the same 89 recorded iterations, interventions modify held-out raw text *before full re-encoding through frozen BERT*; they do not remove or replace cached token representations. Eligible texts come from the 290 specified sequences in the species-balanced lexical panel described above. One agent prepares each edit and another independently reviews it before model outcomes are inspected, checking attribute preservation, repeated evidence, cross-references and grammatical repairs. The resulting edits are agent-reviewed, not human-validated.

Cue deletion and its control. Figure 13 evaluates deletion of each cue family. Coarse-cue deletion removes all evidence for upper-tail and throat colors; fine-grained-cue deletion removes head-pattern and leg-color evidence. The other cue family and unrelated context are preserved, with grammatical repair where needed. To control for generic shortening and editing effects, each deletion is paired with a noncue deletion that preserves all attribute evidence and removes a BERT-token count within 20% of the cue deletion. Sources without an accepted matched control are excluded from the adjusted estimate; eligibility is fixed before outcomes.

Let $D_f(x)$ be deletion of cue family f from source x, and $N_f(x)$ its matched noncue control. For a response m_u, the plotted adjusted loss is

$$L_f^{\mathrm{adj}}(x,u) = \big[m_u(x) - m_u(D_f(x))\big] - \big[m_u(x) - m_u(N_f(x))\big] = m_u(N_f(x)) - m_u(D_f(x)).$$

Positive values indicate a larger loss after cue deletion than after unrelated-text deletion. With source species y and its coarse-compatible species group G_y, the responses are coarse-group probability $P_u(G_y \mid x)$, full-label probability $P_u(y \mid x)$, and conditional fine-grained gain

$$g_u(x) = \frac{P_u(y \mid x)/P_u(G_y \mid x) - 1/|G_y|}{1 - 1/|G_y|}.$$

Gain contrasts require $|G_y| > 1$ and group probability above 10^{-8} for original, edited and control texts jointly. The reference y, G_y stays fixed after editing; negative values are retained. No additional denominator exclusions occur in the plotted data.

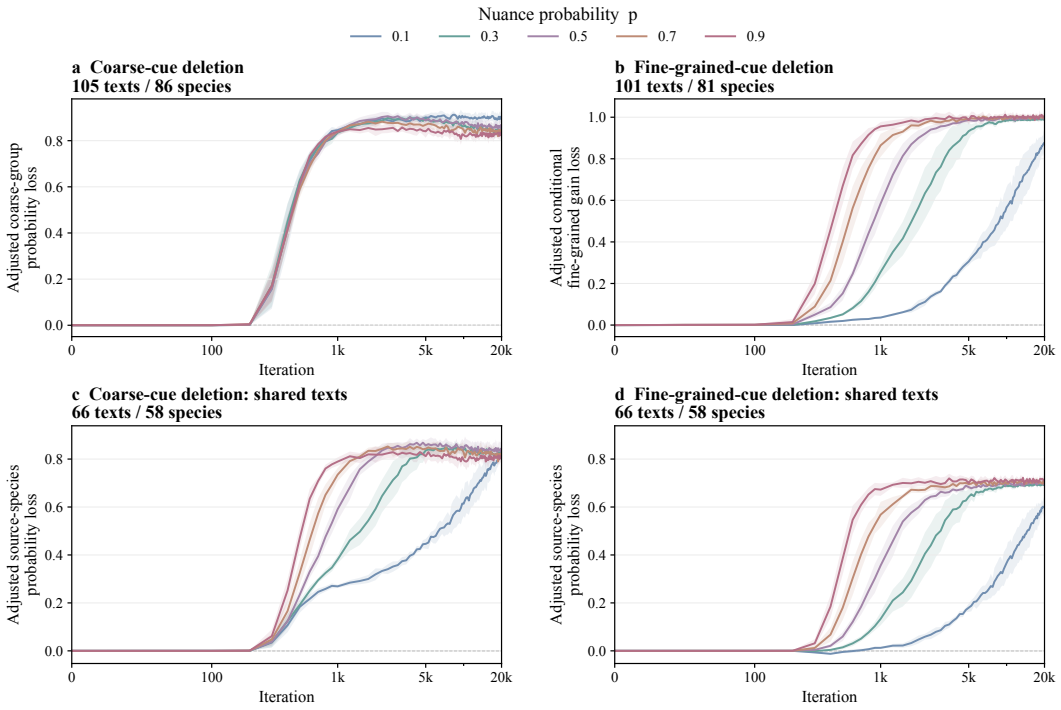


Figure 13: Section C.2.2: cue deletion, adjusted for matched noncue deletion. (a) Coarse-group probability loss on 105 texts from 86 species. (b) Conditional fine-grained gain loss on 101 texts from 81 species. (c,d) Full-label probability losses after coarse and fine-grained deletion on the same 66 texts from 58 species. Only the bottom panels share both response units and population for a direct comparison. Curves and bands show seven-seed means and sample standard deviations after equal-species averaging. All 89 recorded iterations are shown; the symmetric-log axis is linear through iteration 100. Effects are unsmoothed and unclipped.

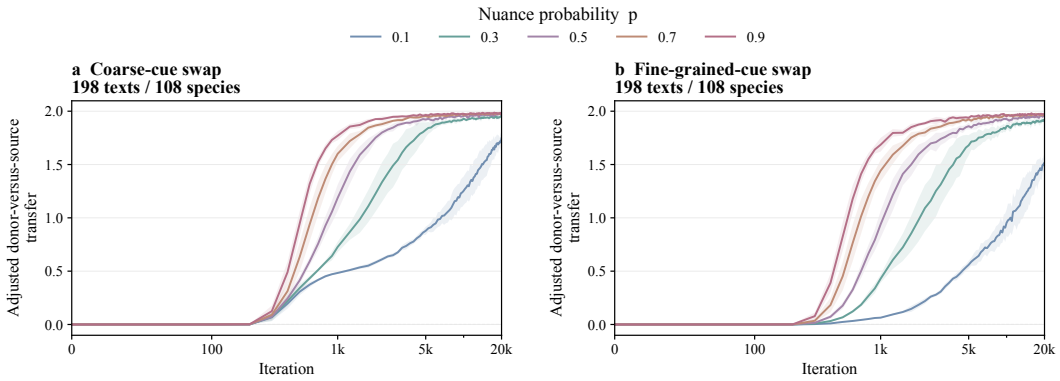


Figure 14: Section C.2.2: donor-versus-source transfer after (a) coarse-cue and (b) fine-grained-cue swaps, adjusted for same-value rewrites. Both panels use the same 198 specified texts from 108 species, with 387 coarse-swap and 378 fine-grained-swap donor variants. Donors are averaged within each source before sources within species and species equally; curves and bands then show seven-seed means and sample standard deviations. The shared vertical scale is a difference of probability differences, not a probability. All 89 iterations are retained with symmetric-log spacing, linear through iteration 100.

Table 2: Section C.2.2: an accepted cue-intervention example. Each swap yields an existing species profile while preserving the other cue family. These are excerpts of edited test inputs; training passages are unchanged.

Condition	Attribute-bearing sentence
Original	Its upper tail is brown, its throat is grey, its head is capped, and its legs are buff.
Coarse deletion	Its head is capped, and its legs are buff.
Fine-grained deletion	Its upper tail is brown and its throat is grey.
Coarse swap	Its upper tail is black, its throat is red, its head is capped, and its legs are buff.
Fine-grained swap	Its upper tail is brown, its throat is grey, its head is unmarked, and its legs are black.
Coarse rewrite	Its upper tail is brown in color, its throat is grey in color, its head is capped, and its legs are buff.
Fine-grained rewrite	Its upper tail is brown, its throat is grey, its head is capped, and its legs are buff in color.

Cue swapping and its control. Figure 14 evaluates whether predictions follow swapped cues. A coarse swap replaces coarse values with those of a donor species while preserving the source's fine-grained values; a fine-grained swap does the reverse. The resulting four-attribute tuple must uniquely identify the donor among the 145 retained species. We edit the source wording rather than transplanting a donor passage. Up to two distinct donors per source and cue family are assigned deterministically from attribute metadata, with donor-use balancing, before outcomes. Each swap is paired with a same-value rewrite of that cue family, which changes wording but preserves the original attributes.

For donor d, let $S_{f,d}(x)$ be the swap, $W_f(x)$ the rewrite, and $v_{d,u}(z) = P_u(d \mid z) - P_u(y \mid z)$. The plotted transfer is

$$R^{\mathrm{adj}}_{f,d}(x,u) = \big[v_{d,u}(S_{f,d}(x)) - v_{d,u}(x)\big] - \big[v_{d,u}(W_f(x)) - v_{d,u}(x)\big].$$

The same donor is used in both contrasts, even when two donors share one rewrite. Positive transfer indicates a shift toward the donor beyond the rewrite response. It lies in $[-2, 2]$ and is not a probability.

Concrete example. Table 2 gives exact sentence excerpts from one accepted source and its edits, included in both figures' shared populations. Unquoted context is unchanged. Both matched noncue controls instead remove "Note each feature calmly before the tour starts." This removes nine BERT tokens versus eleven for either cue deletion, an 18.2% mismatch. References to studying a body part remain when they convey no attribute value.

Aggregation and interpretation. We form each controlled contrast within the same source, model and iteration, then average donors within source, sources within species, and species equally within a run. Finally, each of the seven seeds contributes equally to the mean and sample standard deviation. Population membership is constant across all runs and iterations. At low p, coarse-cue effects emerge before strong fine-grained-cue effects; the latter develop earlier as p increases. For example, at $p = 0.1$ and iteration 20,000, the shared-population fine-grained deletion loss is 0.599 and swap transfer is 1.515. These interventions measure whole-model sensitivity to edited inputs. They do not establish mediation by attention, and small early effects do not establish zero cue use. Matched controls reduce editing confounds; the reviewed subsets do not represent every test passage.

C.3 BIRDTEXT GENERATION WITH GPT-2 MODEL

Experiment C.3 uses $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and paired seeds $12545, 12645, 12745$, giving fifteen runs of $60{,}000$ iterations. Each run uses $100{,}000$ distinct training passages, with pN specified and $(1-p)N$ general sequences. The model is pretrained GPT-2 (124M), with frozen base weights and its tied language-model head. LoRA uses rank 256, scaling parameter $\alpha = 512$ and dropout 0.1, targeting attention `c_attn`/`c_proj` and MLP `c_fc`/`c_proj` modules. AdamW uses learning rate 5×10^{-4}, weight decay 0.01, batch size 32 and gradient clipping at 1, with BF16 arithmetic, 3000 warmup iterations and cosine decay through iteration $60{,}000$. The causal language-model objective supervises the real terminal end-of-text (EOT) token once per example and excludes padding by sequence length. Held-out language-model loss uses the same terminal-inclusive mask and is aggregated by token count; it is distinct from the candidate cross-entropy plotted below.

Candidate scoring and generation. All measurements use the same 2500 specified and 2500 general test passages. Candidate scoring sums the conditional log probabilities of a species name's tokens, excluding EOT, and predicts the highest-scoring of 145 supplied names. Scores are normalized across those names to obtain candidate probabilities. Figure 15 reports the corresponding true-species cross-entropy, pooled over the test passages and separately by sequence type; the balanced pooled population is independent of training p. These probabilities are candidate-normalized, not unrestricted generation probabilities.

Generation is unconstrained greedy decoding, stopped by EOT or any decoded line break, with a cap of 40 new tokens and no species-name-based stopping. An answer maps to a species only when its entire normalized content matches an established species surface form. Coarse-label and full-label accuracy use the species-based definitions in Section C.2; unmapped answers remain in the evaluation denominator rather than being discarded.

The 25 post-initial evaluation iterations are

$$100, 150, 250, 400, 600, 800, 1000, 1200, 1400, 1650, 1950, 2300, 2700,$$
$$3150, 3750, 4400, 5200, 6100, 7200, 8500, 10{,}000, 15{,}000, 25{,}000, 40{,}000, 60{,}000.$$

A single shared iteration-zero baseline is displayed across runs; it is not three independent initial observations. Figures 15, 16 and 17 retain every scheduled measurement, individual seed curves, arithmetic means and one sample-standard-deviation band across three runs at each nuance probability. The iteration axis uses $\log(1+u)$ spacing. Curves are neither smoothed nor aligned to learning events.

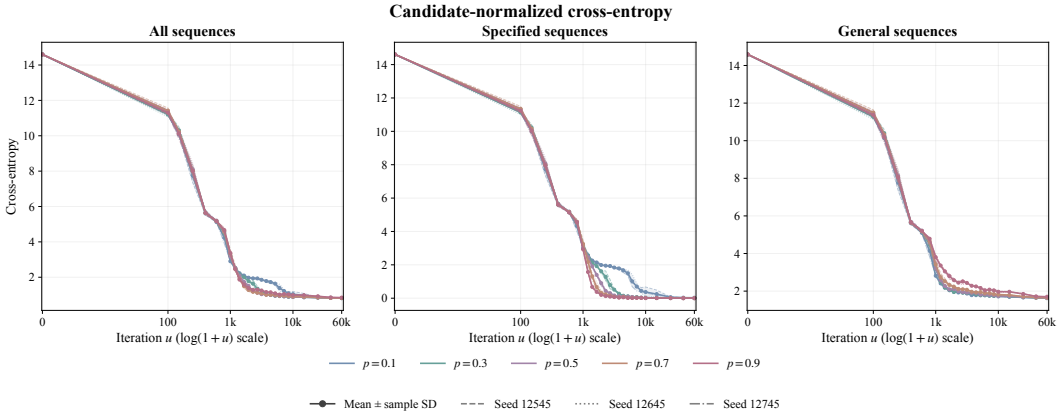


Figure 15: Experiment C.3: candidate-normalized true-species cross-entropy over all 5000 test passages (left), the 2500 specified passages (middle), and the 2500 general passages (right). Markers denote saved observations; bands summarize three seeds.

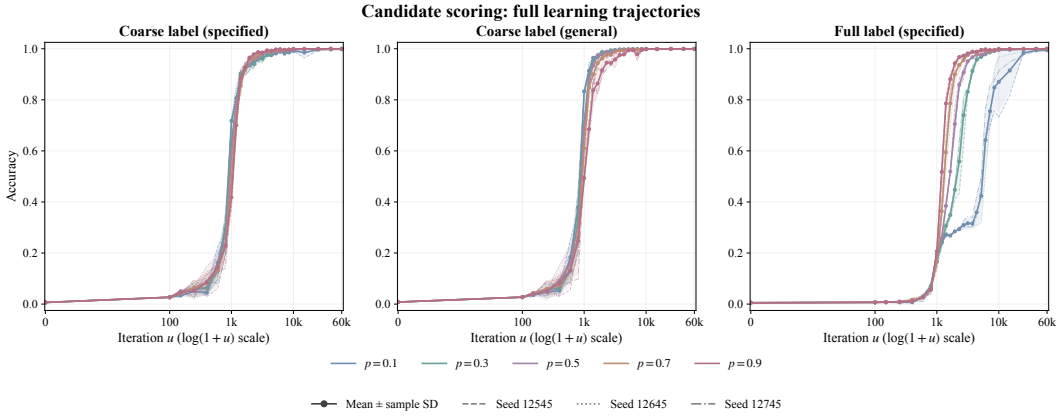


Figure 16: Experiment C.3: candidate-scoring coarse accuracy on specified and general sequences, and full-label accuracy on specified sequences, over all 60,000 iterations. The three seed trajectories, means and sample-standard-deviation bands are retained.

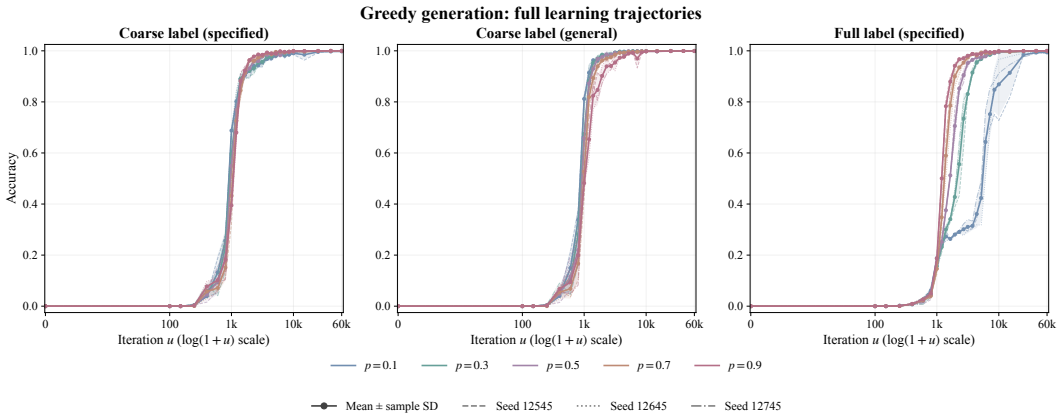


Figure 17: Experiment C.3: greedy-generation coarse accuracy on specified and general sequences, and full-label accuracy on specified sequences, using the same runs, test populations and saved observations as Figure 16. Unmapped complete answers are included in the accuracy denominators.

D PRELIMINARIES AND THE MEANING OF THE CONSTANTS

Throughout, $\eta = \sigma^2/d$ denotes the variance of the noise in a unit direction. The component laws $\mathcal{D}_g$ and $\mathcal{D}_s$ mean the two sampling procedures in Section 2, including when the mixture weight is $p = 0$ or $p = 1$. Thus no conditioning on a probability-zero event is needed at these endpoints.

Write

$$\dot{\boldsymbol{W}} = -\nabla_{\boldsymbol{W}}\mathcal{L}, \qquad \dot{\boldsymbol{v}} = -\nabla_{\boldsymbol{v}}\mathcal{L}, \qquad \boldsymbol{W}(0) = \boldsymbol{0}, \quad \boldsymbol{v}(0) = \boldsymbol{0}. \tag{20}$$

The population vector field is smooth and its solution exists for every finite t. Indeed, putting $R(\boldsymbol{X}) = \max_t \|\boldsymbol{x}_t\|$, differentiation of the two softmax maps gives

$$\|\nabla_{\boldsymbol{W}}\ell_{\mathrm{CE}}\|_F \le 2R(\boldsymbol{X}), \qquad \|\nabla_{\boldsymbol{v}}\ell_{\mathrm{CE}}\| \le 4\|\boldsymbol{W}\|_F R(\boldsymbol{X})^2.$$

All moments of $R(\boldsymbol{X})$ are finite. Consequently $\|\boldsymbol{W}(t)\|_F \le Ct$ and $\|\boldsymbol{v}(t)\| \le C(t + t^2)$ on finite intervals; local uniqueness therefore extends globally. The same moment bounds justify differentiating all the expectations below. Along the flow,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\boldsymbol{W}(t), \boldsymbol{v}(t)) = -\|\nabla_{\boldsymbol{W}}\mathcal{L}\|_F^2 - \|\nabla_{\boldsymbol{v}}\mathcal{L}\|^2. \tag{21}$$

Except in the preceding global-existence bounds, where constants may depend on the fixed data law, the letters c, C denote positive constants depending only on K, L, M, T and on any fixed accuracy or margin thresholds explicitly under discussion. They may change between displays. A statement $o(1)$ in the small-p proof is uniform over $0 \le \eta \le p/(\log(1/p))^\gamma$ and over the time interval specified there. Thus each such statement means a bound by a deterministic function of p tending to zero, not a pointwise-in-time assertion. No constant depends separately on d or σ.

E PROOF OF THEOREM 1: THE REDUCED ODE

Strategy. First identify the parameter subspace fixed by the label and pattern permutations. The common-row shift of the prediction head removes its label-independent part. A common attention-score shift removes one query coordinate in the noiseless case; a reflection removes the same coordinate under Gaussian noise. Finally compute the Gram matrix of the four remaining coordinate directions. Gradient flow in these non-orthonormal coordinates is the inverse-Gram-matrix flow in the theorem.

E.1 THE NOISELESS CASE $\sigma = 0$

Let $n = K + L + M$ and $\boldsymbol{r} = \boldsymbol{\mu} + \boldsymbol{\nu} + \boldsymbol{\xi}$. All clean tokens belong to $\mathrm{span}(\mathcal{V})$. The sample gradients therefore show that every row of $\boldsymbol{W}(t)$ and $\boldsymbol{v}(t)$ belongs to this span.

Permuting the signal patterns and the corresponding coarse labels leaves the loss unchanged. The same holds for the nuance patterns and fine labels, and for the non-relevant patterns. Uniqueness of (20), whose initial point is fixed by these transformations, implies that the trajectory is fixed by them. In particular,

$$\boldsymbol{v} = \bar{v}_\mu\boldsymbol{\mu} + \bar{v}_\nu\boldsymbol{\nu} + \bar{v}_\xi\boldsymbol{\xi}.$$

For clarity, the fixed-point assertion for the head can be checked row by row. The stabilizer of (k, l) permutes all $\boldsymbol{\mu}_i$ with $i \ne k$, all $\boldsymbol{\nu}_j$ with $j \ne l$, and all $\boldsymbol{\xi}_m$. Hence, for scalar coefficients independent of (k, l),

$$\boldsymbol{W}_{(k,l),:}^\top = a\boldsymbol{\mu}_k + b\boldsymbol{\mu} + c\boldsymbol{\nu}_l + e\boldsymbol{\nu} + f\boldsymbol{\xi}.$$

Here these five temporary coefficients only describe the fixed-point space. Since the sample cross-entropy gradient has zero sum over output rows,

$$\boldsymbol{1}_{KL}^\top\dot{\boldsymbol{W}} = \boldsymbol{0}^\top, \qquad \boldsymbol{1}_{KL}^\top\boldsymbol{W} = \boldsymbol{0}^\top.$$

Summing the preceding row formula gives $b = -a/K$, $e = -c/L$, and $f = 0$. Thus,

$$\boldsymbol{W} = w_\mu\boldsymbol{S} + w_\nu\boldsymbol{N}. \tag{22}$$

The row directions are $\boldsymbol{S}_{(k,l),:} = (\boldsymbol{\mu}_k - \boldsymbol{\mu}/K)^\top$ and $\boldsymbol{N}_{(k,l),:} = (\boldsymbol{\nu}_l - \boldsymbol{\nu}/L)^\top$. In particular $\boldsymbol{W}\boldsymbol{r} = \boldsymbol{0}$.

Every clean token has inner product one with $\boldsymbol{r}$. Replacing $\boldsymbol{v}$ by $\boldsymbol{v} + s\boldsymbol{r}$ therefore adds the same scalar to every attention score and leaves the loss unchanged. It follows that $\langle \nabla_{\boldsymbol{v}} \mathcal{L}, \boldsymbol{r} \rangle = 0$. Since $\boldsymbol{v}(0) = \boldsymbol{0}$,

$$K\bar{v}_\mu + L\bar{v}_\nu + M\bar{v}_\xi = 0.$$

Using $v_\mu = \bar{v}_\mu - \bar{v}_\xi$ and $v_\nu = \bar{v}_\nu - \bar{v}_\xi$ gives exactly

$$\boldsymbol{v} = v_\mu \boldsymbol{\mu} + v_\nu \boldsymbol{\nu} - \frac{K v_\mu + L v_\nu}{n} \boldsymbol{r}. \tag{23}$$

E.2 The noisy case $\sigma > 0$

Recall that $n = K + L + M$ and $\boldsymbol{r} = \boldsymbol{\mu} + \boldsymbol{\nu} + \boldsymbol{\xi}$, so $\|\boldsymbol{r}\|^2 = n$. The permutation argument from Section E.1 remains valid because isotropic Gaussian noise is invariant under orthogonal transformations. Moreover, orthogonal transformations that act identically on $\mathrm{span}(\mathcal{V})$ and arbitrarily on its orthogonal complement preserve the data law. Uniqueness of the zero-initialized gradient flow therefore keeps every row of $\boldsymbol{W}$ and the query $\boldsymbol{v}$ in $\mathrm{span}(\mathcal{V})$. The same permutation fixed-point calculation, together with the zero sum of the head rows, gives

$$\boldsymbol{W} = w_\mu \boldsymbol{S} + w_\nu \boldsymbol{N}, \qquad \boldsymbol{v} = \bar{v}_\mu \boldsymbol{\mu} + \bar{v}_\nu \boldsymbol{\nu} + \bar{v}_\xi \boldsymbol{\xi}.$$

In particular, $\boldsymbol{W}\boldsymbol{r} = \boldsymbol{0}$ throughout the flow. It remains to show that the query component along $\boldsymbol{r}$ stays zero. For a noisy token $\boldsymbol{x}_t = \boldsymbol{x}_t^{\mathrm{clean}} + \sigma \boldsymbol{\varepsilon}_t / \sqrt{d}$,

$$\langle \boldsymbol{x}_t, \boldsymbol{v} + s\boldsymbol{r} \rangle - \langle \boldsymbol{x}_t, \boldsymbol{v} \rangle = s + \frac{s\sigma}{\sqrt{d}} \langle \boldsymbol{\varepsilon}_t, \boldsymbol{r} \rangle.$$

When $\sigma > 0$, the additional noise term generally differs across tokens, so shifting $\boldsymbol{v}$ along $\boldsymbol{r}$ need not preserve the attention weights. Instead, a reflection symmetry of the population loss will imply that its derivative along $\boldsymbol{r}$ vanishes whenever $\boldsymbol{v} \perp \boldsymbol{r}$.

A reflection preserving the noisy data law. Define

$$\boldsymbol{R} = \boldsymbol{I} - \frac{2}{n} \boldsymbol{r}\boldsymbol{r}^\top.$$

Since $\|\boldsymbol{r}\|^2 = n$, we have $\boldsymbol{R}^\top = \boldsymbol{R}$, $\boldsymbol{R}^2 = \boldsymbol{I}$, and $\boldsymbol{R}\boldsymbol{r} = -\boldsymbol{r}$: $\boldsymbol{R}$ reflects vectors across $\boldsymbol{r}^\perp$. Every clean pattern $\boldsymbol{u} \in \mathcal{V}$ satisfies $\langle \boldsymbol{u}, \boldsymbol{r} \rangle = 1$, and hence $\boldsymbol{R}\boldsymbol{u} = \boldsymbol{u} - 2\boldsymbol{r}/n$. Thus the affine map $\boldsymbol{x} \mapsto \boldsymbol{R}\boldsymbol{x} + 2\boldsymbol{r}/n$ fixes every clean pattern; geometrically, it is reflection across the affine hyperplane $\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{r} \rangle = 1\}$ containing all clean tokens. Define

$$\boldsymbol{X}' = \boldsymbol{R}\boldsymbol{X} + \frac{2}{n} \boldsymbol{r} \boldsymbol{1}_T^\top. \tag{24}$$

Its columns satisfy

$$\boldsymbol{x}_t' = \boldsymbol{x}_t^{\mathrm{clean}} + \frac{\sigma}{\sqrt{d}} \boldsymbol{R}\boldsymbol{\varepsilon}_t.$$

Because $\boldsymbol{R}$ is orthogonal, the vectors $\boldsymbol{R}\boldsymbol{\varepsilon}_t$ are again independent standard Gaussians. Consequently, $\boldsymbol{X}'$ and $\boldsymbol{X}$ have the same conditional law given the clean token arrangement and the label. This transformation reverses the noise component along $\boldsymbol{r}$ while leaving the clean sequence and label unchanged.

Reflective symmetry of the population loss. Fix a head satisfying $\boldsymbol{W}\boldsymbol{r} = \boldsymbol{0}$. Then $\boldsymbol{W}\boldsymbol{R} = \boldsymbol{W}$, and for any query $\boldsymbol{v}$,

$$\boldsymbol{W}\boldsymbol{X}' = \boldsymbol{W}\boldsymbol{X}, \qquad (\boldsymbol{X}')^\top \boldsymbol{R}\boldsymbol{v} = \boldsymbol{X}^\top \boldsymbol{v} - \frac{2}{n} (\boldsymbol{r}^\top \boldsymbol{v}) \boldsymbol{1}_T.$$

The second identity shows that reflecting both the noisy input and the query changes all attention scores by the same scalar. Softmax removes this common shift, so

$$f_{\boldsymbol{W}, \boldsymbol{R}\boldsymbol{v}}(\boldsymbol{X}') = f_{\boldsymbol{W}, \boldsymbol{v}}(\boldsymbol{X}).$$

Taking expectations and using the equality of the conditional data laws gives

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{R}\boldsymbol{v}) = \mathcal{L}(\boldsymbol{W}, \boldsymbol{v}) \qquad \text{whenever } \boldsymbol{W}\boldsymbol{r} = \boldsymbol{0}.$$

This is a population-loss symmetry obtained by pairing reflected noise realizations; it does not assert that replacing $\boldsymbol{v}$ by $\boldsymbol{R}\boldsymbol{v}$ preserves the loss on each fixed noisy input.

Preservation of the query constraint. Write $\boldsymbol{v} = \boldsymbol{v}_\perp + a\boldsymbol{r}$, where $\boldsymbol{r}^\top \boldsymbol{v}_\perp = 0$. Since $\boldsymbol{R}\boldsymbol{v} = \boldsymbol{v}_\perp - a\boldsymbol{r}$, the preceding identity becomes

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{v}_\perp + a\boldsymbol{r}) = \mathcal{L}(\boldsymbol{W}, \boldsymbol{v}_\perp - a\boldsymbol{r}).$$

Thus the loss is an even function of a, although it need not be constant in a. Differentiating at $a = 0$ yields

$$0 = \frac{d}{da}\mathcal{L}(\boldsymbol{W}, \boldsymbol{v}_\perp + a\boldsymbol{r})\Big|_{a=0} = \boldsymbol{r}^\top \nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{v}_\perp).$$

Therefore, whenever $\boldsymbol{r}^\top \boldsymbol{v} = 0$, the query gradient is tangent to $\boldsymbol{r}^\perp$. The head subspace has already been shown invariant, and $\boldsymbol{v}(0) = \boldsymbol{0}$ belongs to $\boldsymbol{r}^\perp$. Uniqueness of the joint gradient flow now implies

$$\boldsymbol{r}^\top \boldsymbol{v}(t) = 0 \qquad \text{for every } t \geq 0.$$

Using the three-coordinate representation of $\boldsymbol{v}$, this gives

$$K\bar{v}_\mu + L\bar{v}_\nu + M\bar{v}_\xi = 0.$$

With $v_\mu = \bar{v}_\mu - \bar{v}_\xi$ and $v_\nu = \bar{v}_\nu - \bar{v}_\xi$, we recover

$$\boldsymbol{v} = v_\mu \boldsymbol{\mu} + v_\nu \boldsymbol{\nu} - \frac{Kv_\mu + Lv_\nu}{n}\boldsymbol{r},$$

which is exactly (23). Hence the parameterizations (22)–(23) remain valid for $\sigma > 0$: the noiseless shift symmetry is replaced by reflection symmetry, which suffices to preserve the same subspace from zero initialization.

E.3 The coordinate metric, in both cases

The two head directions are orthogonal and satisfy

$$\|\boldsymbol{S}\|_F^2 = L(K-1), \qquad \|\boldsymbol{N}\|_F^2 = K(L-1), \qquad \langle \boldsymbol{S}, \boldsymbol{N} \rangle_F = 0.$$

The two query coordinate directions are $\boldsymbol{\mu} - K\boldsymbol{r}/n$ and $\boldsymbol{\nu} - L\boldsymbol{r}/n$. Their Gram matrix and its inverse are

$$\boldsymbol{G}_v = \begin{bmatrix} K - K^2/n & -KL/n \\ -KL/n & L - L^2/n \end{bmatrix}, \qquad \boldsymbol{G}_v^{-1} = \begin{bmatrix} 1/K + 1/M & 1/M \\ 1/M & 1/L + 1/M \end{bmatrix}.$$

Let $\boldsymbol{\theta} = (w_\mu, w_\nu, v_\mu, v_\nu)^\top$, and let $\boldsymbol{G}$ be the block-diagonal Gram matrix with head block $\mathrm{diag}(L(K-1), K(L-1))$ and query block $\boldsymbol{G}_v$. The gradient is tangent to the subspace just proved invariant. The chain rule then gives

$$\dot{\boldsymbol{\theta}} = -\boldsymbol{G}^{-1} \nabla \mathcal{L}_s(\boldsymbol{\theta}), \qquad \boldsymbol{\theta}(0) = \boldsymbol{0}. \tag{25}$$

This is precisely the four-scalar ODE in Theorem 1. Its image under (22)–(23) solves (20); uniqueness identifies the two solutions. This proves Theorem 1 in both cases. □

We will use the equivalent form of loss dissipation

$$-\dot{\mathcal{L}}_s = \nabla \mathcal{L}_s^\top \boldsymbol{G}^{-1} \nabla \mathcal{L}_s. \tag{26}$$

Since K, L, M are fixed, this quadratic form and the squared Euclidean norm of the four coordinate derivatives differ by fixed positive factors.

F Conditional Gaussian identities

These identities will be used for all three noise regimes. They also make explicit why the effective noise parameter is $\eta = \sigma^2/d$, rather than σ itself.

Given the attention vector $\boldsymbol{a} = \mathrm{Softmax}(\boldsymbol{X}^\top \boldsymbol{v})$, put

$$a = \boldsymbol{a}_{t_s}, \qquad b = \begin{cases} 0, & z = 0, \\ \boldsymbol{a}_{t_n}, & z = 1, \end{cases} \qquad A_2 = \sum_{t=1}^{T} \boldsymbol{a}_t^2.$$

Thus a and b are the signal and nuance attention weights, and A_2 is the variance multiplier for the averaged token noise. These symbols in this section refer to random, noisy attention weights.

The query in (23) is orthogonal to every signal contrast $\boldsymbol{\mu}_i - \boldsymbol{\mu}/K$ and every nuance contrast $\boldsymbol{\nu}_j - \boldsymbol{\nu}/L$. Gaussian noise in these mutually orthogonal subspaces is independent. Therefore, conditional on the attention weights and the clean token arrangement, the coarse logits, up to a common shift, have the law

$$w_\mu\big(a\,\boldsymbol{e}_k + \sqrt{\eta A_2}\,\boldsymbol{g}_\mu\big), \qquad \boldsymbol{g}_\mu \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K),$$

and the fine logits have the law

$$w_\nu\big(b\,\boldsymbol{e}_l + \sqrt{\eta A_2}\,\boldsymbol{g}_\nu\big), \qquad \boldsymbol{g}_\nu \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_L).$$

Here $\boldsymbol{e}_k$ and $\boldsymbol{e}_l$ are the appropriate standard basis vectors, and the two Gaussian vectors are independent. The attention law is independent of the identities of k and l, since all patterns within a group have the same query projection. Centering these Gaussian vectors would produce the actual contrast logits; centering does not change softmax.

For $n \geq 2$, define the scalar conditional loss and conditional correct-class probability

$$\Phi_n(w, a, A_2) = \mathbb{E}_{\boldsymbol{g}} \log \sum_{j=1}^{n} \exp\big(w(a\,\mathbf{1}_{\{j=1\}} + \sqrt{\eta A_2}\,g_j)\big) - wa, \tag{27}$$

$$P_n(w, a, A_2) = \mathbb{E}_{\boldsymbol{g}} \frac{\exp\big(w(a + \sqrt{\eta A_2}\,g_1)\big)}{\sum_{j=1}^{n} \exp\big(w(a\,\mathbf{1}_{\{j=1\}} + \sqrt{\eta A_2}\,g_j)\big)}. \tag{28}$$

Each output logit is the sum of a term depending only on its coarse index and a term depending only on its fine index. Thus the joint softmax factors into its coarse and fine softmax distributions, and the sample cross-entropy is the sum of their cross-entropies. Averaging the conditional Gaussian laws gives the exact identity

$$\begin{aligned} \mathcal{L}_s = (1-p)\mathbb{E}_g\big[\Phi_K(w_\mu, a, A_2) + \Phi_L(w_\nu, 0, A_2)\big] \\ + p\mathbb{E}_s\big[\Phi_K(w_\mu, a, A_2) + \Phi_L(w_\nu, b, A_2)\big]. \end{aligned} \tag{29}$$

Here the outer expectations average over attention and the clean token arrangement; the contrast noise has already been integrated inside each Φ_n. The notation $\mathcal{L}_s$ denotes the reduced *population* loss, not the specified-component loss. In particular, general examples supply a fine-head term with weight $1-p$, whereas the fine-label signal has weight p. When $\eta = 0$, the former term is exactly $(1-p)\log L$.

The corresponding accuracies satisfy

$$Q_{\mu,g} = \mathbb{E}_g P_K(w_\mu, a, A_2), \qquad Q_{\mu,s} = \mathbb{E}_s P_K(w_\mu, a, A_2), \tag{30}$$

$$Q_\nu = \mathbb{E}_s\big[P_K(w_\mu, a, A_2) P_L(w_\nu, b, A_2)\big]. \tag{31}$$

Conditional independence of the two Gaussian logit vectors is what justifies the product in (31).

Lemma F.1 (Nonnegative heads and conditional accuracy bounds)**.** *For every $p \in [0,1]$, $\eta \geq 0$, and $t \geq 0$, both head scalars are nonnegative. For $w, a \geq 0$,*

$$\frac{1}{n} \leq P_n(w, a, A_2) \leq \frac{1}{n} + \frac{wa}{4}. \tag{32}$$

If $\eta > 0$, both head scalars also satisfy

$$0 \leq w_\mu(t), w_\nu(t) \leq \frac{1+\sqrt{T}}{2\eta}. \tag{33}$$

Proof. At $w = 0$, $\partial_w \Phi_n = -(n-1)a/n \leq 0$. The attention does not depend on the head scalars. The head components of (25) thus point into the nonnegative quadrant at its boundary, proving its invariance.

At $a = 0$, exchangeability gives $P_n = 1/n$. Differentiating the true-class shift gives

$$\partial_a P_n = w\mathbb{E}[q_1(1-q_1)] \in [0, w/4].$$

Integrating from zero to a proves (32).

For the upper head bound, condition on a, A_2. In the Gaussian experiment $\boldsymbol{u} = a\boldsymbol{e}_j + \sqrt{\eta A_2}\boldsymbol{g}$ with a uniform class j, the posterior is $\mathrm{Softmax}((a/(\eta A_2))\boldsymbol{u})$. Thus the conditional expected cross-entropy $\Phi_n(w, a, A_2)$, which is convex in w, is minimized at $w = a/(\eta A_2)$. This also follows directly by inserting the Gaussian densities and canceling their common factors. Its derivative is consequently nonnegative for $w \geq a/(\eta A_2)$.

If one coordinate of an attention vector equals a, Cauchy–Schwarz gives

$$A_2 \geq a^2 + \frac{(1-a)^2}{T-1}.$$

The maximum of $a/(a^2 + (1-a)^2/(T-1))$ over $0 \leq a \leq 1$ is $(1+\sqrt{T})/2$: differentiation gives the maximizing point $a = 1/\sqrt{T}$. This bound applies equally to the nuance coordinate. The head vector field points inward at the upper boundary in (33), proving the claim. □

A useful exact derivative follows from Gaussian integration by parts:

$$\partial_w \Phi_n = -a(1 - P_n) + \eta A_2 w\, \mathbb{E}\Big[1 - \sum_{j=1}^{n} q_j^2\Big]. \tag{34}$$

For example, the Gaussian contribution equals $\sqrt{\eta A_2}\sum_j \mathbb{E}[g_j q_j] = \eta A_2 w \sum_j \mathbb{E}[q_j(1-q_j)]$.

Finally introduce the excess population loss above the unavoidable general-sequence fine-label loss:

$$\mathcal{E}(t) = \mathcal{L}(\boldsymbol{W}(t), \boldsymbol{v}(t)) - (1-p)\log L. \tag{35}$$

On a general sequence the fine label is independent and uniform. Its expected cross-entropy is at least $\log L$, by Jensen's inequality. All other cross-entropy terms are nonnegative. Therefore

$$\begin{aligned} &\mathcal{E}(t) \geq 0, \qquad \dot{\mathcal{E}}(t) \leq 0, \\ &(1-p)\mathbb{E}_g[-\log q_\mu(k)] \leq \mathcal{E}(t), \qquad p\mathbb{E}_s[-\log q_{(k,l)}] \leq \mathcal{E}(t), \end{aligned} \tag{36}$$

where $q_\mu(k) = \sum_l q_{(k,l)}$. In particular, for $0 < p \leq 1/2$,

$$\mathcal{E}(t) \leq \delta p \quad \Longrightarrow \quad Q_{\mu,g}(t) \geq 1 - 2\delta p, \quad Q_{\mu,s}(t), Q_\nu(t) \geq 1 - \delta. \tag{37}$$

We used $1 - x \leq -\log x$ and $q_{(k,l)} \leq q_\mu(k)$. This sublevel set is invariant for all future time, a fact that will establish the "for every $t \geq \tau_{\nu,1}$" clause of Theorem 2 without any claim of monotonicity of individual accuracies.

G Proof of Theorem 2: the noiseless case

Strategy and intuition. We first identify the bottleneck on the signal-only trajectory ($p = 0$). Its signal head grows like $\log t$, while the signal query coordinate grows like twice the logarithm of that head. The query metric simultaneously increases the nuance coordinate, but by the smaller factor $K/(K+M)$. Consequently a specified sequence receives nuance attention of order $(\log t)^{-\gamma}$, where $\gamma = 2M/(K+M)$. This suppresses the fine-label driving force beyond its mixture factor p.

We compare the small-p trajectory with this reference only through a controlled warm-up ending at $t_* = p^{-\alpha}$. Thereafter we bound the actual trajectory directly, up to a prescribed fine-label margin and a fixed multiple of $(\log(1/p))^\gamma/p$. A query combination cancels the force from general-sequence signal learning and isolates the feedback between the nuance head and its attention. This feedback forces a first fixed positive margin at time of order $(\log(1/p))^\gamma/p$.

The transition uses a different argument. At fixed positive fine-label margins the nuance head has size of order $\sqrt{\log\log(1/p)}$, while nuance attention still vanishes: the head compensates for small attention. A definite loss decrease prevents the margin from returning near zero, and loss dissipation forces entry into a small excess-loss sublevel within order $1/p$. An upper margin-speed bound gives the matching lower transition time; no monotonicity of the margin is assumed. Finally, the exact invariant sublevel (37) gives success for every later time, without extending the parameter bounds indefinitely.

G.1 EXACT SCALAR EQUATIONS

For $n \geq 2$ and $u \geq 0$, write

$$\ell_n(u) = \log(1 + (n-1)e^{-u}), \qquad r_n(u) = \frac{n-1}{e^u + n - 1}, \qquad \pi_n(u) = 1 - r_n(u). \tag{38}$$

Thus $\ell_n'(u) = -r_n(u)$; $\pi_n(u)$ is the correct-class softmax probability for a true-class margin u.

The following attention weights and margins always refer to the *clean* tokens evaluated at the current parameters:

$$a_g = \frac{e^{v_\mu}}{e^{v_\mu} + T - 1}, \quad a_s = \frac{e^{v_\mu}}{e^{v_\mu} + e^{v_\nu} + T - 2}, \quad b_s = \frac{e^{v_\nu}}{e^{v_\mu} + e^{v_\nu} + T - 2}, \tag{39}$$

$$c_s = 1 - a_s - b_s, \quad u_g = w_\mu a_g, \quad u_s = w_\mu a_s, \quad u_\nu = w_\nu b_s. \tag{40}$$

Here c_s is the total non-relevant attention in a specified sequence; it is zero if $T = 2$. Put $D_\mu = L(K-1)$ and $D_\nu = K(L-1)$, the squared norms of the two head directions. When $\eta = 0$, the loss is exactly

$$\mathcal{L}_s = (1-p)\log L + (1-p)\ell_K(u_g) + p\ell_K(u_s) + p\ell_L(u_\nu). \tag{41}$$

Consequently

$$D_\mu \dot{w}_\mu = (1-p)a_g r_K(u_g) + p a_s r_K(u_s), \tag{42}$$

$$D_\nu \dot{w}_\nu = p b_s r_L(u_\nu), \tag{43}$$

$$-\partial_{v_\mu}\mathcal{L}_s = (1-p)w_\mu a_g(1-a_g)r_K(u_g) + p w_\mu a_s(1-a_s)r_K(u_s) - p u_\nu a_s r_L(u_\nu), \tag{44}$$

$$-\partial_{v_\nu}\mathcal{L}_s = p u_\nu (1-b_s) r_L(u_\nu) - p w_\mu a_s b_s r_K(u_s). \tag{45}$$

The query velocity is $-\boldsymbol{G}_v^{-1}$ times the two query derivatives. In particular both heads are nondecreasing in this case. Both sequence types contribute to the same coarse head in (42), so its driving force has no overall factor p. By contrast, the fine-head driving force in (43) contains both the mixture factor p and the nuance attention b_s. The fine-label part of (45) likewise contains $p w_\nu b_s(1-b_s)$. These are the two sources of suppression that the incubation argument will quantify.

Two combinations of the existing query coordinates will be useful:

$$\rho = \frac{K}{K+M}, \qquad \chi = v_\nu - \rho v_\mu, \qquad \mathcal{R} = K v_\mu + L v_\nu. \tag{46}$$

The scalar χ measures nuance-query growth beyond that induced by signal learning alone. The scalar $\mathcal{R} = -(K+L+M)\bar{v}_\xi$ measures the common separation of the relevant groups from the non-relevant group. The two driving forces

$$G = (1-p)w_\mu a_g(1-a_g)r_K(u_g), \qquad J = p u_\nu r_L(u_\nu) \tag{47}$$

are, respectively, the general-sequence signal-attention force and the specified-sequence nuance-attention force. Direct substitution in the query equations gives

$$\dot{v}_\mu = \left(\frac{1}{K} + \frac{1}{M}\right) G - \left(\frac{a_s}{K} - \frac{c_s}{M}\right) J + e_\mu, \tag{48}$$

$$\dot{v}_\nu = \frac{G}{M} + \left(\frac{1-b_s}{L} + \frac{c_s}{M}\right) J + e_\nu, \tag{49}$$

$$\dot{\chi} = \left(\frac{1}{L} + \frac{1}{K+M}\right)\left((1-b_s)J - p w_\mu a_s b_s r_K(u_s)\right), \tag{50}$$

$$\dot{\mathcal{R}} = \frac{K+L+M}{M}\left(G + p w_\mu a_s c_s r_K(u_s) + c_s J\right) \geq 0. \tag{51}$$

The temporary error terms in (48)–(49) contain *only* the specified-sequence coarse-loss derivatives; in particular

$$|e_\mu| + |e_\nu| \leq C p w_\mu r_K(u_s). \tag{52}$$

They have not discarded any fine-label force. The cancellation of G in (50) is essential: v_ν itself can increase through the off-diagonal query metric even when $p = 0$, whereas χ removes precisely that signal-induced motion.

G.2 THE SIGNAL-ONLY TRAJECTORY AND A CONTROLLED WARM-UP

Use a superscript 0 for the noiseless $p = 0$ trajectory. Then $w_\nu^0 = 0$ and $v_\nu^0 = \rho v_\mu^0$. Dividing the two nonzero signal equations gives

$$\frac{dv_\mu^0}{dw_\mu^0} = \left(\frac{1}{K} + \frac{1}{M}\right) D_\mu w_\mu^0 \frac{T-1}{e^{v_\mu^0} + T - 1}.$$

Integration from zero yields the exact identity

$$e^{v_\mu^0} + (T-1)v_\mu^0 = 1 + \frac{(T-1)D_\mu}{2}\left(\frac{1}{K} + \frac{1}{M}\right)(w_\mu^0)^2. \tag{53}$$

The head increases without bound: a finite limiting head would, by this identity, give a finite query and a strictly positive limiting head derivative. Equation (53) therefore implies

$$e^{v_\mu^0} \asymp (w_\mu^0)^2, \qquad a_g^0 = 1 - O((w_\mu^0)^{-2}).$$

Inserting these relations into the head equation gives $\dot{w}_\mu^0 \asymp e^{-w_\mu^0}$ for large time. Integrating the derivative of $e^{w_\mu^0}$ proves

$$w_\mu^0 = \log(1+t) + O(1), \qquad v_\mu^0 = 2\log w_\mu^0 + O(1), \qquad b_s^0 \asymp (w_\mu^0)^{-\gamma}, \quad \gamma = 2(1-\rho) = \frac{2M}{K+M}. \tag{54}$$

The last relation follows directly from $b_s^0/a_s^0 = \exp(-(1-\rho)v_\mu^0)$ and $a_s^0 \to 1$. Both clean coarse margins on this reference trajectory tend to infinity. Moreover a_g^0 and a_s^0 are increasing functions of $v_\mu^0 \geq 0$, so these margins are increasing.

Lemma G.1 (Warm-up)**.** *There is an exponent $\alpha \in (0,1)$, depending only on K, L, M, T, such that, with*

$$H = \log(1/p), \qquad t_* = p^{-\alpha},$$

with $p < e^{-e}$, the noiseless $p > 0$ trajectory is $o(1)$-close to the $p = 0$ trajectory uniformly on $[0, t_*]$. The error remains $o(1)$ after multiplication by any fixed power of H. At t_*,

$$\begin{aligned} &w_\mu = \alpha H + O(1), \qquad v_\mu = 2\log H + O(1), \qquad w_\nu = o(1), \quad \chi = o(1), \\ &\mathcal{R} = 2(K + L\rho)\log H + O(1). \end{aligned} \tag{55}$$

Proof. Here is a quantitative continuity argument, so that continuity is not being used on an unspecified growing time interval. In a fixed-radius tube around $\boldsymbol{\theta}^0(t)$, the Jacobian of its $p = 0$ vector field is bounded by $C/(1+t)$. To check this bound, the only nonconstant loss term is $\ell_K(w_\mu a_g)$; for large t,

$$a_g' = O((w_\mu^0)^{-2}), \quad a_g'' = O((w_\mu^0)^{-2}), \quad |\ell_K'| + |\ell_K''| \leq Ce^{-w_\mu^0}.$$

The same bounds hold up to constants throughout the fixed-radius tube. Its loss Hessian, and hence its vector-field Jacobian, is at most $Ce^{-w_\mu^0} \leq C/(1+t)$. For bounded t, enlarge C.

On this tube the change in the vector field from replacing zero by p is at most $Cp(1+\log(2+t))^m$ for a fixed integer m: the extra derivatives in (42)–(45) have this polynomial bound. Gronwall's inequality therefore gives, until the first exit from the tube,

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^0(t)\| \leq Cp(1+t)^{C+1}(1+\log(2+t))^m.$$

Choose $\alpha < 1/(2(C+2))$. The right side, even multiplied by any fixed power of H, tends to zero uniformly for $t \leq p^{-\alpha}$. It is therefore smaller than the tube radius, which rules out a first exit. Equation (55) now follows from (54) and (53). □

G.3 BOUNDS DURING THE INCUBATION AND TRANSITION INTERVALS

For a fixed margin bound $B > 0$, define

$$\tau_B = \inf\{t \geq t_* : u_\nu(t) = B\}, \qquad \inf \varnothing = \infty. \tag{56}$$

The following continuation estimates keep both coarse margins large, control the feedback between nuance-head and nuance-query growth, and convert that feedback into bounds on nuance attention. They are stated on a stopped interval: their proof does not assume that τ_B is finite.

Lemma G.2 (Continuation up to a fixed fine-label margin)**.** *For every fixed $B, D > 0$, the noiseless trajectory, for sufficiently small p, satisfies the following bounds on*

$$t_* \le t \le \min\{\tau_B, DH^\gamma/p\} :$$

$$cH \le w_\mu \le CH, \qquad \kappa \log H - C \le v_\mu \le 2\log H + C, \quad \kappa = \frac{2(K+L\rho)}{K+L} > 0, \tag{57}$$

$$-o(1) \le \chi \le \gamma \log H + C, \qquad w_\nu^2 \le C(\chi+1), \qquad \chi \le C(w_\nu^2+1), \tag{58}$$

$$a_g = 1 - o(1), \qquad a_s = 1 - o(1), \qquad b_s = o(1). \tag{59}$$

In particular $w_\nu^2 \le C\log H$ and $\|\boldsymbol{v}\| \le C\log H$. On any fixed positive margin band $0 < b_- \le u_\nu \le B$ within this interval, one also has

$$c\log H \le w_\nu^2 \le C\log H. \tag{60}$$

The leading constants can be chosen independently of the fixed horizon multiplier D; the smallness threshold for p may depend on D.

Proof. Stop additionally at the first violation of the loose bounds below. Starting from (55), we strictly improve those bounds before any such exit, which proves the estimates on the entire stated interval.

Start with the loose bounds $w_\mu \in [\alpha H/2, 10H]$, $v_\mu \in [(\kappa/2)\log H, C\log H]$, $a_s \ge 1/3$, $b_s \le 2/3$, and sufficiently large fixed upper bounds of order $\sqrt{\log H}$ and $\log H$ on w_ν and $|\chi|$. They hold at t_*. Under these loose bounds, (52) and the specified coarse term in (50) have size at most

$$CpHe^{-cH}, \qquad \int_{t_*}^{DH^\gamma/p} CpHe^{-cH}\,dt = o(1). \tag{61}$$

First, (50) implies $\chi \ge -o(1)$. Since $(w_\nu^2)' = 2J/D_\nu$, and $1 - b_s \in [1/3, 1]$, integrating the same equation gives both inequalities in (58) that relate χ to w_ν^2.

Second, the coefficient $a_s/K - c_s/M$ in (48) is positive for large H, since $a_s/c_s = e^{v_\mu}/(T-2)$ when $T > 2$. Also

$$G \le D_\mu w_\mu (1-a_g)\dot{w}_\mu.$$

Multiplying (48) by $e^{v_\mu} + T - 1$, integrating, and using (61) yields

$$e^{v_\mu} + (T-1)v_\mu \le e^{v_\mu(t_*)} + (T-1)v_\mu(t_*) + \frac{D_\mu(T-1)}{2}\left(\frac{1}{K} + \frac{1}{M}\right)(w_\mu^2 - w_\mu(t_*)^2) + o(1) \le CH^2.$$

The error is still $o(1)$ after multiplication by $e^{v_\mu} + T - 1$, because the loose upper bound makes that multiplier a fixed power of H. This proves the upper bound on v_μ.

Third, $\mathcal{R}$ is nondecreasing. If $w_\nu \le 2B + 1$, then $\chi \le C$, and $\mathcal{R} = (K + L\rho)v_\mu + L\chi$ proves $v_\mu \ge 2\log H - C$. If $w_\nu > 2B + 1$, then $b_s = u_\nu/w_\nu < 1/2$. Since $a_s \ge 1/3$,

$$\mathcal{R} = (K+L)v_\mu + L\log(b_s/a_s) \le (K+L)v_\mu + L\log(3/2),$$

which proves $v_\mu \ge \kappa \log H - C$. Moreover

$$\chi = \log(b_s/a_s) + (1-\rho)v_\mu \le \log 3 + (1-\rho)v_\mu \le \gamma \log H + C.$$

This also proves $w_\nu^2 \le C\log H$.

Fourth, $b_s = o(1)$ uniformly. Indeed, if $w_\nu \le (\log H)^{1/4}$, the bound $\chi \le C(1 + w_\nu^2)$ gives

$$\frac{b_s}{a_s} = e^{\chi - (1-\rho)v_\mu} \le C\exp(C\sqrt{\log H})H^{-(1-\rho)\kappa} = o(1).$$

For larger w_ν, $b_s \le B/(\log H)^{1/4} = o(1)$. Also $c_s \le (T-2)e^{-v_\mu} = o(1)$ and $1 - a_g \le (T-1)e^{-v_\mu} = o(1)$. This proves (59) and strictly improves the loose attention bounds.

Finally w_μ is nondecreasing. For its upper bound, whenever $w_\mu \ge 9H$, the loose attention bounds and (42) give $\dot{w}_\mu \le Ce^{-w_\mu/3} \le Ce^{-3H}$. Over the entire allowed horizon the resulting increase is $o(1)$, so the head cannot rise from $9H$ to $10H$.

All the loose inequalities have now been strictly improved by taking their fixed constants sufficiently large and p sufficiently small. The leading constants can be chosen independently of the fixed horizon multiplier D; only the required smallness of p depends on D. No first exit is possible. This proves (57)–(59).

It remains to prove the lower head bound on a fixed band $0 < b_- \le u_\nu \le B$. Use

$$\chi = \log u_\nu - \log w_\nu - \log a_s + (1-\rho)v_\mu \ge (1-\rho)\kappa \log H - \log w_\nu - C.$$

Since $w_\nu \ge u_\nu \ge b_-$ and $w_\nu^2 \le C \log H$, we have $\log w_\nu = O(\log\log H)$. Combining the preceding inequality with $\chi \le C(1 + w_\nu^2)$ proves (60). □

Classification with small nuance attention. On every fixed band $0 < b_- \le u_\nu \le B$ in the stopped interval, the identity $u_\nu = w_\nu b_s$ now gives

$$w_\nu \asymp \sqrt{\log H}, \qquad b_s \asymp (\log H)^{-1/2}. \tag{62}$$

Thus a fixed positive fine-label margin is attained while nuance attention still vanishes; classification depends on the head–attention product, not attention alone. On this band, head growth contributes only

$$b_s \dot{w}_\nu = \frac{p}{D_\nu} b_s^2 r_L(u_\nu) = O\left(\frac{p}{\log H}\right)$$

to the margin velocity. The positive nuance-query contribution in (71) below is instead of order p. The general-sequence query force can oppose it, so this comparison is not a lower bound on the net velocity $\dot{u}_\nu$. The transition proof uses loss dissipation rather than an assumed monotone margin.

G.4 The first fixed fine-label margin is reached at order H^γ/p

For each fixed $B > 0$, we now prove that the hitting time (56) is finite and

$$c\frac{H^\gamma}{p} \le \tau_B \le C\frac{H^\gamma}{p}. \tag{63}$$

The horizon constant in the preceding subsection is chosen large enough to contain the upper bound obtained below.

For a lower bound, restrict initially to $w_\nu \le 1$ and $\chi \le 1$. Then $\mathcal{R} \ge \mathcal{R}(t_*)$, together with the upper bound in Lemma G.2, gives $v_\mu = 2\log H + O(1)$. Since $\chi \ge -o(1)$,

$$b_s \asymp H^{-\gamma}, \qquad D_\nu \dot{w}_\nu \asymp pH^{-\gamma}, \qquad (1-b_s)J \asymp p w_\nu H^{-\gamma}. \tag{64}$$

This transfers the signal-only attention suppression to an initial portion of the actual small-p trajectory. In particular, (43) and (50) give

$$\dot{w}_\nu \le CpH^{-\gamma}, \qquad \dot{\chi} \le CpH^{-\gamma} w_\nu.$$

Thus, for a sufficiently small fixed $c_0 > 0$, these two scalars remain below $1/2$ on

$$[t_*, s_0], \qquad s_0 = t_* + c_0 H^\gamma/p.$$

In particular $u_\nu \le CH^{-\gamma} = o(1)$ there, so $\tau_B > s_0$.

At this time we also have the useful loss bound

$$\mathcal{E}(s_0) \le p\log L + o(p). \tag{65}$$

Indeed $a_g, a_s \ge 1/3$ and, for $u \ge 0$, $\ell_K(u) \le K r_K(u)$. Therefore

$$-\partial_{w_\mu}\mathcal{L}_s \ge c\big((1-p)\ell_K(u_g) + p\ell_K(u_s)\big) \ge c(\mathcal{E} - p\log L)_+.$$

Loss dissipation implies $\dot{\mathcal{E}} \le -c(\mathcal{E} - p\log L)_+^2$. Integrating this scalar differential inequality over $[t_*, s_0]$ proves $\mathcal{E}(s_0) \le p\log L + C/(s_0 - t_*) = p\log L + o(p)$.

For an upper bound on τ_B, (57) and $a_s \ge 1/3$ imply

$$b_s = a_s e^{\chi - (1-\rho)v_\mu} \ge cH^{-\gamma} e^{\chi}. \tag{66}$$

Since $\chi \geq -o(1)$ and $r_L(u_\nu) \geq r_L(B) > 0$, the head reaches some fixed $w_* > 0$ within CH^γ/p, unless $u_\nu = B$ has already been reached. It cannot subsequently decrease. After that time, (50), (61), and (66) give

$$\dot{\chi} \geq cpH^{-\gamma}e^{\chi}, \qquad \text{hence} \qquad \frac{\mathrm{d}}{\mathrm{d}t}e^{-\chi} \leq -cpH^{-\gamma}.$$

The positive quantity $e^{-\chi}$ cannot obey this inequality for longer than another CH^γ/p. Therefore $u_\nu = B$ must be reached. This proves (63). The constants used in this argument are fixed before taking p small; the first-exit estimates above apply on the resulting finite multiple of H^γ/p.

G.5 A FIXED-MARGIN TRANSITION TAKES ORDER $1/p$

Choose $B_0 \in (0, 1)$ sufficiently small that

$$\pi_L(B_0) - 1/L < \varepsilon/3, \tag{67}$$

and set $\tau_{\nu,0} = \tau_{B_0}$. Define

$$\varepsilon_* = \min\{\varepsilon/4, \ell_L(B_0)/4\} > 0.$$

We will take $\tau_{\nu,1}$ to be the first time after $\tau_{\nu,0}$ when $\mathcal{E} \leq \varepsilon_* p$.

We first show that reaching B_0 dissipates a fixed positive amount of loss after division by p. Let s_1 be the last time before $\tau_{\nu,0}$ with $u_\nu = B_0/2$. Since $u_\nu(s_0) = o(1)$ and $\tau_{\nu,0}$ is the first hit of B_0, we have $s_1 > s_0$ and

$$B_0/2 \leq u_\nu \leq B_0 \quad \text{on } [s_1, \tau_{\nu,0}].$$

The last crossing gives this band condition without any monotonicity assumption. By (60), $w_\nu \geq 1$ on this band for small p, so

$$\nabla u_\nu = (0, b_s, -u_\nu a_s, u_\nu(1 - b_s))^\top, \qquad \|\nabla u_\nu\| \leq C u_\nu.$$

The head-coordinate bound uses $b_s = u_\nu/w_\nu$. In (45), the positive fine-label force is bounded below by cpu_ν, while the specified coarse term is exponentially small. Thus $|\partial_{v_\nu}\mathcal{L}_s| \geq cpu_\nu$. The fixed positive-definite metric in (25)–(26) gives

$$\sqrt{-\dot{\mathcal{E}}} \geq cpu_\nu, \qquad \|\dot{\boldsymbol{\theta}}\| \leq C\sqrt{-\dot{\mathcal{E}}}, \qquad |\dot{u}_\nu| \leq Cu_\nu\sqrt{-\dot{\mathcal{E}}}.$$

Combining the first and last inequalities yields $-\dot{\mathcal{E}} \geq cp|\dot{u}_\nu|$: dissipated loss controls the total variation of the margin. Therefore

$$\begin{aligned} \mathcal{E}(s_1) - \mathcal{E}(\tau_{\nu,0}) &\geq cp \int_{s_1}^{\tau_{\nu,0}} |\dot{u}_\nu(t)|\, dt \\ &\geq cp\big|u_\nu(\tau_{\nu,0}) - u_\nu(s_1)\big| = cpB_0/2. \end{aligned} \tag{68}$$

Using $\mathcal{E}(s_1) \leq \mathcal{E}(s_0)$ and (65), and absorbing its $o(p)$ remainder, proves that for a fixed $\delta_0 > 0$,

$$\mathcal{E}(\tau_{\nu,0}) \leq p(\log L - \delta_0). \tag{69}$$

At every subsequent time in the continued interval, the exact inequality $\mathcal{E} \geq p\ell_L(u_\nu)$ and loss monotonicity imply

$$u_\nu \geq b_- > 0, \tag{70}$$

where b_- depends only on δ_0, L. This loss gap, rather than monotonicity of u_ν, prevents a return to the unlearned state.

Next we record a uniform upper speed for this margin. Differentiating $u_\nu = w_\nu b_s$ and using (48)–(49) gives

$$\dot{u}_\nu = b_s\dot{w}_\nu - u_\nu\left(\frac{a_s}{K} - \frac{c_s}{M}\right)G + pu_\nu^2 r_L(u_\nu)\left(\frac{a_s^2}{K} + \frac{(1-b_s)^2}{L} + \frac{c_s^2}{M}\right) + o(p). \tag{71}$$

The last term is uniform on any fixed bounded margin interval, by (61). On a fixed positive margin band, the coefficient of $pu_\nu^2 r_L(u_\nu)$ tends to $1/K + 1/L$, so this positive query contribution is indeed of order p. The coefficient of G is negative; it is not omitted in the identity. For the upper bound we may drop that nonpositive term. Moreover $b_s\dot{w}_\nu \leq Cp$ and $\sup_{u \geq 0} u^2 r_L(u) < \infty$. Hence, with a constant independent of the chosen finite upper margin bound,

$$\dot{u}_\nu \leq C_1 p \tag{72}$$

for sufficiently small p.

Choose B_1 so that $\ell_L(B_1) \le \varepsilon_*/4$. Before the target loss sublevel is reached, there are two alternatives. If $b_- \le u_\nu \le B_1$, then (45) gives $|\partial_{v_\nu}\mathcal{L}_s| \ge cp$. If $u_\nu > B_1$, then

$$\mathcal{E} = (1-p)\ell_K(u_g) + p\ell_L(u_\nu) + o(p)$$

by (57)–(59); thus $\mathcal{E} > \varepsilon_* p$ implies $(1-p)\ell_K(u_g) \ge \varepsilon_* p/2$. The head derivative then has magnitude at least cp. In both alternatives, loss dissipation yields

$$-\dot{\mathcal{E}} \ge c_1 p^2. \tag{73}$$

These constants only involve b_-, B_1 and the fixed problem parameters, not a further upper bound on u_ν.

The order of choices is important. The speed constant C_1 uses the global bound on $u^2 r_L(u)$, not a bound obtained by replacing u with a prospective $B_{\max}$. The dissipation constant c_1 uses only the fixed band $[b_-, B_1]$ or the fixed coarse-loss threshold $\varepsilon_* p/2$. We can therefore first choose $C_2 > \log L/c_1$, then fix $B_{\max} > B_0 + C_1 C_2 + 1$, and only afterward take p small enough for the estimates with this upper margin. Choose also a fixed horizon multiplier D large enough to contain the already proved upper bound for τ_{B_0} and an additional C_2/p; this is possible since $H^\gamma \to \infty$.

Apply Lemma G.2 with these fixed $B_{\max}, D$. Equation (70) rules out a lower-margin exit, and (72) rules out reaching $B_{\max}$ within C_2/p. If the target sublevel were not reached by then, (73) would dissipate more than the available excess loss $p \log L$. Thus the sublevel is reached and

$$\tau_{\nu,1} - \tau_{\nu,0} \le C_2/p.$$

At $\tau_{\nu,0}$, $\mathcal{E} \ge p\ell_L(B_0) > \varepsilon_* p$, so the two times are distinct. At $\tau_{\nu,1}$, $\ell_L(u_\nu) \le \varepsilon_*$, whereas $\ell_L(B_0) \ge 4\varepsilon_*$. Thus u_ν must have increased by a fixed positive amount. Integrating (72) gives the matching lower bound

$$c/p \le \tau_{\nu,1} - \tau_{\nu,0} \le C/p. \tag{74}$$

All intervals just used fit within a fixed multiple of H^γ/p, since $H^\gamma \to \infty$.

G.6 The accuracy assertions

Choose a fixed τ_ε so that on the reference $p = 0$ trajectory both coarse probabilities and both signal attention weights a_g^0, a_s^0 are at least $1 - \varepsilon/3$ for all $t \ge \tau_\varepsilon$. This is possible by (54) and the monotonicity established there. Lemma G.1 gives these bounds with $1 - 2\varepsilon/3$ on $[\tau_\varepsilon, t_*]$ for small p, and the fine margin is uniformly $o(1)$ there. Thus signal attention is concentrated by the chosen p-independent coarse-learning time and stays so throughout the warm-up portion of the plateau. On $[t_*, \tau_{\nu,0}]$, the coarse margins are at least cH, whereas $0 \le u_\nu \le B_0$. In the noiseless case,

$$Q_{\mu,g} = \pi_K(u_g), \qquad Q_{\mu,s} = \pi_K(u_s), \qquad Q_\nu = \pi_K(u_s)\pi_L(u_\nu).$$

These identities, (67), and the warm-up bounds prove all three hallucination inequalities throughout $[\tau_\varepsilon, \tau_{\nu,0}]$, after decreasing p_ε if necessary.

At $\tau_{\nu,1}$ the invariant sublevel (37) applies with $\delta = \varepsilon_*$. It proves all the eventual-accuracy inequalities for *every* later time. Equations (63) and (74) give the asserted time scales. Finally enlarge or decrease the constants to arrange $0 < c_\varepsilon < 1 < C_\varepsilon$ and $\tau_\varepsilon < \tau_{\nu,0}$. This proves Theorem 2 when $\sigma = 0$. □

H Proof of Theorem 2: positive noise

Assume now

$$0 < \eta \le pH^{-\gamma}, \qquad H = \log(1/p).$$

Strategy. The noiseless proof uses four ingredients: the warm-up, bounds on the parameter path until a fixed fine-label margin is reached, the incubation-time comparison, and loss dissipation during the transition. We establish explicit Gaussian estimates for each. The only noise term that is not negligible in the incubation comparison is a damping term proportional to ηw_ν in the nuance-head equation. It cannot prevent escape, because the nuance query keeps increasing and raises the signal driving that head. The proof below retains this damping term rather than treating noise as a uniform small perturbation over the entire long time interval.

H.1 Quantitative Gaussian estimates

Keep the *clean* weights and margins (39)–(40) as functions of the current noisy-flow parameters. They are reference quantities, not the random noisy attention weights in Section F.

By token permutation invariance, the signal can be placed first and the nuance, when present, second. If $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_T)$, the random attention vector has exactly the law

$$\mathrm{Softmax}\big(\boldsymbol{c}_z + \sqrt{\eta}\,\|\boldsymbol{v}\|\boldsymbol{g}\big), \qquad \boldsymbol{c}_0 = (v_\mu, 0, \dots, 0)^\top, \quad \boldsymbol{c}_1 = (v_\mu, v_\nu, 0, \dots, 0)^\top. \tag{75}$$

It is independent of the two contrast-noise vectors used in (27). Substituting this attention law into (29) is therefore an exact finite-dimensional representation of the noisy population loss; the remaining $d - (K + L + M)$ noise coordinates do not enter it.

For the noisy trajectory, use the same definition (56) of τ_B, now for its clean reference margin u_ν. Fix $B, D > 0$. All continuation and accumulated-error estimates below are stopped at $\min\{\tau_B, DH^\gamma/p\}$, and initially also at the first exit from the following loose region. With fixed sufficiently large loose constants, assume

$$\begin{gathered} \alpha H/2 \le w_\mu \le 10H, \qquad (\kappa/2)\log H \le v_\mu \le C\log H, \qquad |v_\nu| \le C\log H, \\ 0 \le w_\nu \le C\sqrt{\log H}, \qquad a_s \ge 1/3, \quad b_s \le 2/3, \qquad 0 \le u_\nu \le B. \end{gathered} \tag{76}$$

For some fixed integer m and positive constant c, Gaussian differentiation gives

$$D_\mu \dot{w}_\mu = \big((1-p)a_g r_K(u_g) + p a_s r_K(u_s)\big)(1 + o(1)), \tag{77}$$

$$D_\nu \dot{w}_\nu = p b_s r_L(u_\nu)(1 + o(1)) - \frac{L-1}{L}\eta w_\nu (1 + o(1)), \tag{78}$$

$$\dot{v}_\mu = \left(\frac{1}{K} + \frac{1}{M}\right) G - \left(\frac{a_s}{K} - \frac{c_s}{M}\right) J + e_\mu, \tag{79}$$

$$\dot{v}_\nu = \frac{G}{M} + \left(\frac{1-b_s}{L} + \frac{c_s}{M}\right) J + e_\nu, \tag{80}$$

$$|e_\mu| + |e_\nu| \le C\left(\eta H^m G + pHe^{-cH} + p\eta H^m + \eta w_\nu^2 e^{-v_\mu}\right). \tag{81}$$

The two relative $o(1)$ terms in (78) need not be equal. The estimate means that each is bounded uniformly in absolute value by a quantity tending to zero; no cancellation between them is assumed. Also, uniformly on (76),

$$\mathcal{E} = (1-p)\ell_K(u_g) + p\ell_L(u_\nu) + o(p), \tag{82}$$

$$\partial_{v_\nu}\mathcal{L}_s = -(1-b_s)J + o(pH^{-\gamma}), \tag{83}$$

$$Q_{\mu,g} = \pi_K(u_g) + o(1), \qquad Q_{\mu,s} = \pi_K(u_s) + o(1), \qquad Q_\nu = \pi_K(u_s)\pi_L(u_\nu) + o(1). \tag{84}$$

The required error budget has three parts. Query errors of size $o(pH^{-\gamma})$ accumulate to $o(1)$ over a fixed multiple of H^γ/p. The head damping ηw_ν need not be negligible on this scale and is retained. Finally, the signal-query estimate multiplies its equation by $e^{v_\mu} + T - 1$; its accumulated error need only be $o(H^2)$, not $o(1)$. The general-sequence fine-loss query derivative requires particular care: its extra factor e^{-v_μ} comes from attention concentration near the signal vertex. A bound of merely $O(\eta\,\mathrm{poly}(H))$ would not suffice. We derive these estimates next.

Gaussian differentiation identities. For a smooth function F and a standard Gaussian vector,

$$\frac{d}{ds}\mathbb{E}F(\boldsymbol{c} + \sqrt{s}\boldsymbol{g}) = \frac{1}{2}\mathbb{E}\Delta F(\boldsymbol{c} + \sqrt{s}\boldsymbol{g}). \tag{85}$$

This follows by differentiating for $s > 0$ and using $\mathbb{E}[g_j h(\boldsymbol{g})] = \mathbb{E}[\partial_j h(\boldsymbol{g})]$; continuity gives the identity at zero. In (75), the corresponding query derivative is

$$\partial_{v_i}\mathbb{E}F = \mathbb{E}\big[(\partial_{v_i}\boldsymbol{c}_z)^\top \nabla F\big] + \frac{\eta}{2}\,\partial_{v_i}\|\boldsymbol{v}\|^2\,\mathbb{E}\Delta F, \qquad i \in \{\mu, \nu\}. \tag{86}$$

All expectations in this display evaluate F at the random argument in (75).

For the classifier noise, the log-sum-exp Hessian has trace $1 - \sum_j q_j^2 \le 1$. Hence

$$0 \le \Phi_n(w, a, A_2) - \ell_n(wa) \le \frac{1}{2}\eta w^2 A_2. \tag{87}$$

For a true-class margin wa, the sharper estimate

$$0 \le \Phi_n(w, a, A_2) - \ell_n(wa) \le C\eta w^2 e^{-wa} e^{C\eta w^2} \tag{88}$$

follows by using $1 - \sum_j q_j^2 \le 2(1 - q_1)$ and $1 - q_1 \le \sum_{j \ne 1} e^{s_j - s_1}$ at logit vector $\boldsymbol{s}$ in (85). The Gaussian exponential moments are explicit. Differentiating these same expressions gives the analogous derivative bounds with fixed polynomial factors in $1 + w$.

For completeness, the relative attention-smoothing estimates used for the coarse terms do not require a Taylor remainder at an uncontrolled Gaussian argument. For each of

$$f = \ell_K(wa), \qquad f = ar_K(wa), \qquad f = wa(1-a)r_K(wa),$$

where a is the signal softmax coordinate, direct differentiation in the score variables yields $|\partial_i \partial_j f| \le C(1 + w)^4 f$. The same statement holds with $wa(1 - a)$ replaced by $1 - a$ when needed for variance derivatives. Applying (85) and integrating the resulting upper and lower differential inequalities bounds the smoothed expectation between $e^{-Cs(1+w)^4} f$ and $e^{Cs(1+w)^4} f$. Here $s = \eta\|\boldsymbol{v}\|^2$, so these factors are $1 + O(\eta H^m)$ on (76). Combined with (88) and (34), this proves (77), the relative $O(\eta H^m G)$ error in the general coarse query derivative, and the absolute bound $CpHe^{-cH}$ for the specified coarse query derivatives. In the general law, the clean score vector has no v_ν dependence, so its coarse v_ν derivative consists solely of the second term of (86) and has the same relative bound.

Taking f to be any one softmax coordinate instead gives $|\partial_i \partial_j f| \le Cf$. For any query with $\eta\|\boldsymbol{v}\|^2 \le 1$, not only on the loose region, integrating the same heat-identity bounds therefore yields

$$\mathbb{E}_g a = a_g\big(1 + O(\eta\|\boldsymbol{v}\|^2)\big), \quad \mathbb{E}_s a = a_s\big(1 + O(\eta\|\boldsymbol{v}\|^2)\big), \quad \mathbb{E}_s b = b_s\big(1 + O(\eta\|\boldsymbol{v}\|^2)\big). \tag{89}$$

These estimates compare mean noisy attention with clean reference attention; they do not identify individual noisy attention realizations with the clean weights.

The general-sequence fine term. At zero fine signal, Gaussian symmetry and Taylor's formula at the uniform softmax give, uniformly for $\eta w^2 \le 1$,

$$\Phi_L(w, 0, A_2) = \log L + \frac{L-1}{2L}\eta w^2 A_2 + O(\eta^2 w^4), \tag{90}$$

$$\partial_w \Phi_L(w, 0, A_2) = \frac{L-1}{L}\eta w A_2 + O(\eta^2 w^3), \tag{91}$$

$$\partial_{A_2} \Phi_L(w, 0, A_2) = \frac{L-1}{2L}\eta w^2 + O(\eta^2 w^4). \tag{92}$$

The remainders follow from bounded fourth derivatives of log-sum-exp and the finite fourth Gaussian moment; applying the same argument to its derivatives gives the last two displays.

Under the general-sequence attention law, let $\widehat{a}$ be the random signal attention weight. The elementary bound

$$1 - \widehat{a} \le \sum_{t \ne t_s} \exp(s_t - s_{t_s})$$

for score vector $\boldsymbol{s}$ gives

$$\mathbb{E}(1 - \widehat{a}) \le (T-1)e^{-v_\mu} e^{\eta\|\boldsymbol{v}\|^2}. \tag{93}$$

To apply (86), we need derivatives of the entire conditional fine loss as a function of scores. For fixed w_ν, set

$$F(\boldsymbol{s}) = \Phi_L\big(w_\nu, 0, A_2(\mathrm{Softmax}(\boldsymbol{s}))\big).$$

The first two derivatives with respect to the variance multiplier satisfy

$$|\partial_{A_2} \Phi_L| \le C\eta w_\nu^2, \qquad |\partial_{A_2}^2 \Phi_L| \le C\eta^2 w_\nu^4. \tag{94}$$

Indeed, apply (85) to log-sum-exp with variance $\eta w_\nu^2 A_2$. The first derivative is $\eta w_\nu^2/2$ times its expected Hessian trace; the second is $(\eta w_\nu^2)^2/4$ times its expected iterated Laplacian. The fourth derivatives of log-sum-exp are bounded for fixed L, since they are fixed polynomials in softmax probabilities. Thus (94) follows directly, without differentiating an unspecified remainder.

Every fixed-order score derivative of $A_2 = \sum_t \boldsymbol{a}_t^2$ is bounded by $C(1 - \widehat{a})$. For the first derivative, $\partial_{s_i} A_2 = 2\boldsymbol{a}_i(\boldsymbol{a}_i - A_2)$. Subsequent derivatives are fixed polynomials in the attention coordinates that

vanish at the signal vertex, hence are bounded by its distance $2(1-\widehat{a})$ in the ℓ^1 norm. In particular, the chain rule gives

$$\partial_i F = (\partial_{A_2}\Phi_L)\partial_i A_2, \qquad \partial_{ii} F = (\partial^2_{A_2}\Phi_L)(\partial_i A_2)^2 + (\partial_{A_2}\Phi_L)\partial_{ii} A_2.$$

Since $\eta w_\nu^2 \le 1$ on (76) for small p, these formulas imply

$$|\partial_i F| + |\Delta F| \le C\eta w_\nu^2 (1-\widehat{a}).$$

The clean general score vector is $(v_\mu, 0, \dots, 0)^\top$, and $|\partial_{v_i}\|\boldsymbol{v}\|^2| \le C\|\boldsymbol{v}\|$. Equations (86) and (93) consequently give

$$\begin{aligned} |\partial_{v_i}\mathbb{E}_g\Phi_L(w_\nu, 0, A_2)| &\le C\eta w_\nu^2(1+\eta\|\boldsymbol{v}\|)e^{-v_\mu}e^{\eta\|\boldsymbol{v}\|^2} \\ &\le C\eta w_\nu^2 e^{-v_\mu}, \qquad i \in \{\mu,\nu\}. \end{aligned} \tag{95}$$

The absorbed factors tend to one uniformly on (76). Also $\mathbb{E}_g A_2 = 1 + O(e^{-v_\mu})$, so (91) supplies the damping term in (78).

The specified-sequence fine term and the accumulated errors. For bounded u_ν, score derivatives of $b_s r_L(w_\nu b_s)$ are bounded by that same positive quantity times a fixed polynomial in $1 + w_\nu$. The heat identity therefore changes its expectation by a relative $O(\eta H^m)$. Equation (34) adds a variance derivative of size $O(\eta w_\nu)$, but this specified-sequence contribution has the extra mixture factor p. The general fine damping coefficient is consequently multiplied by $1 + o(1)$, proving (78). Its query derivatives differ from the clean derivatives by at most $Cp\eta H^m$, by (86) and the bounded derivatives of the two softmax maps. This proves (81).

On the loose region, $G \le CHe^{-cH}$, $w_\nu^2 \le C\log H$, and $v_\mu \ge (\kappa/2)\log H$. The four contributions in (81), after division by $pH^{-\gamma}$, are bounded respectively by

$$CH^{m+1}e^{-cH}, \qquad CH^{\gamma+1}e^{-cH}, \qquad C\eta H^{m+\gamma}, \qquad C(\log H)H^{-\kappa/2}.$$

Each tends to zero uniformly for $0 < \eta \le pH^{-\gamma}$. Hence

$$|e_\mu| + |e_\nu| = o(pH^{-\gamma}), \qquad \int_{t_*}^{t} (|e_\mu| + |e_\nu|)\, ds = o(1), \tag{96}$$

for $t \le \min\{\tau_B, DH^\gamma/p\}$ as long as the loose region has not been exited. The integral bound uses its length at most DH^γ/p. Multiplying the query-velocity errors by the fixed query metric recovers the corresponding loss-gradient errors; in particular this proves (83).

After multiplying the signal-query equation by $e^{v_\mu} + T - 1$, the last error in (81) instead contributes at most

$$C\int_{t_*}^{t} \eta w_\nu^2\, ds \le CD\log H = o(H^2). \tag{97}$$

This is the comparison required for the signal-query bound. It need not be $o(1)$. The other three weighted errors still integrate to $o(1)$, since the loose upper bound on v_μ adds only a fixed power of H to the preceding vanishing bounds. By contrast, when w_ν is of constant order during early incubation, the damping ηw_ν can be comparable to the positive force $pb_s r_L(u_\nu)$. It is not included among the negligible query errors.

Finally, the general fine-loss excess is $O(\eta w_\nu^2) = o(p)$. The error in the general coarse loss is bounded by $C\eta H^m e^{-cH} = o(p)$, the specified coarse loss is $O(pe^{-cH})$, and the specified fine-loss approximation error is $O(p\eta H^m) = o(p)$. These statements prove (82). For the accuracies, the expected change of the logits from query noise and contrast noise is at most $C\sqrt{\eta}\, H(1+\|\boldsymbol{v}\|) = o(1)$. The softmax map has a globally bounded derivative, and conditional factorization is given by (30)–(31). This proves (84).

H.2 Warm-up and first-exit bounds with noise

The warm-up argument is quantitative also under noise. On the fixed-radius tube around $\boldsymbol{\theta}^0(t)$ used in Lemma G.1, Gaussian centering and Taylor's formula give

$$\|\nabla\mathcal{L}_s^{p,\eta} - \nabla\mathcal{L}_s^{0,0}\| \le C(p+\eta)(1+\|\boldsymbol{\theta}\|)^m.$$

To check this estimate, differentiate the finite-dimensional Gaussian representation (75) and (27); all derivatives of the softmax maps are bounded, and each remaining factor is a polynomial in the parameters and the Gaussian coordinates. The first-order centered Gaussian term has expectation zero, and the second-order remainder is bounded by its finite Gaussian moments. Since $\eta \le p$, the Gronwall proof of Lemma G.1 is unchanged. Thus (55) and the uniform warm-up accuracy approximation remain valid.

We now prove the noisy counterpart of Lemma G.2 on $[t_*, \min\{\tau_B, DH^\gamma/p\}]$, rather than assuming long-time closeness to the noiseless flow. Stop additionally at the first violation of (76); the estimates just derived apply up to that stop. From (79)–(80),

$$\dot\chi = \left(\frac{1}{L} + \frac{1}{K+M}\right)(1-b_s)J + e_\chi, \tag{98}$$

$$\dot{\mathcal{R}} = \frac{K+L+M}{M}(G + c_s J) + e_{\mathcal{R}}, \tag{99}$$

where both new error terms satisfy (96). In particular

$$\chi \ge -o(1), \qquad \mathcal{R}(t) \ge \mathcal{R}(t_*) - o(1). \tag{100}$$

The negative term in (78) gives $w_\nu \dot w_\nu \le CJ$. Combining this with (98) and integrating proves

$$w_\nu^2 \le C(\chi + 1). \tag{101}$$

The upper signal-query estimate is still valid. The fine term in (79) is nonpositive, and (77) gives

$$G \le (1 + o(1)) D_\mu w_\mu (1 - a_g)\dot w_\mu.$$

Multiplying the query equation by $e^{v_\mu} + T - 1$ and integrating gives

$$e^{v_\mu} + (T-1)v_\mu \le CH^2 + C\int_{t_*}^{t} \eta w_\nu^2\, ds + o(1) \le CH^2 + CD\log H \le C'H^2.$$

This uses precisely the weighted budget (97). For each fixed D, the term $CD\log H$ is negligible relative to H^2, so the leading constant C' can be fixed independently of D by taking p sufficiently small afterward. Thus $v_\mu \le 2\log H + C$ without discarding accumulated noise.

The reverse comparison $\chi \le C(1 + w_\nu^2)$ needs a different argument, since w_ν need not be monotone. We supply an inward-pointing barrier. The upper bound on v_μ and $a_s \ge 1/3$ give $b_s \ge cH^{-\gamma}e^{\chi}$. Choose fixed large A_0, A_1 and consider the boundary

$$\chi = A_0 + A_1(w_\nu + 1)^2.$$

At this boundary the signal term in (78) exceeds twice the damping term: indeed it is at least $cpH^{-\gamma}e^{\chi}$, whereas the damping is at most $CpH^{-\gamma}w_\nu$. Increasing A_0 makes this comparison hold for every $w_\nu \ge 0$. Thus $\dot w_\nu \ge cpb_s r_L(u_\nu)$ at the boundary. At the same time,

$$\dot\chi \le Cpw_\nu b_s r_L(u_\nu) + o(pH^{-\gamma}).$$

Taking A_1 sufficiently large yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\chi - A_1(w_\nu+1)^2\big) < 0$$

there. The small error is absorbed because $b_s \ge cH^{-\gamma}$ on this boundary. The trajectory starts strictly below it at t_*. Therefore

$$\chi \le A_0 + A_1(w_\nu+1)^2 \le C(1 + w_\nu^2) \tag{102}$$

throughout the allowed interval.

The remaining first-exit steps now follow from explicit inequalities already established. The lower bound (100), together with (102), gives $v_\mu \ge \kappa\log H - C$ by the two cases $w_\nu \le 2B+1$ and $w_\nu > 2B+1$ used in Section G.3. The identity $\chi = \log(b_s/a_s) + (1-\rho)v_\mu$ gives $\chi \le \gamma\log H + C$ and then $w_\nu^2 \le C\log H$ by (101). Splitting at $w_\nu = (\log H)^{1/4}$ proves $b_s = o(1)$ exactly as before, and $v_\mu \to \infty$ gives $a_g, a_s = 1 - o(1)$. Equation (77) is positive and, for $w_\mu \ge 9H$, is at most Ce^{-3H} up to a factor tending to one. It preserves both head bounds.

All loose first-exit inequalities are again strictly improved, so no additional exit is possible. Hence (57)–(59), and the fixed-band consequence (60), hold throughout the stated stopped interval for the noisy trajectory. Their leading constants can be fixed independently of D; the required smallness of p may depend on a subsequently fixed D.

In particular, on a fixed positive reference-margin band, (62) still holds for w_ν, b_s. Equation (89) gives the same order $(\log H)^{-1/2}$ for *mean* noisy nuance attention. This relates the head–attention compensation to the actual noisy model, without making a pointwise assertion about individual attention realizations.

H.3 Incubation with the damping term retained

As in the noiseless case, on the initial region $w_\nu \le 1$, $\chi \le 1$, the bounds on $\mathcal{R}, v_\mu$ give $v_\mu = 2\log H + O(1)$ and $b_s \asymp H^{-\gamma}$. Thus the positive fine-head force has size $pH^{-\gamma}$, and (89) gives the same suppression for mean noisy attention. This is a statement about the positive force, not the net head velocity, which also contains damping. Equations (78) and (98) give

$$\dot{w}_\nu \le CpH^{-\gamma}, \qquad \dot{\chi} \le CpH^{-\gamma} w_\nu + o(pH^{-\gamma}).$$

Thus the small fixed interval multiplier c_0 can again be chosen so that $u_\nu = o(1)$ until $s_0 = t_* + c_0 H^\gamma / p$. This proves the lower bound on every fixed-margin hitting time.

On this interval w_ν is bounded. Conditional on any attention vector, (87) and $w_\nu b \ge 0$ give a fine-label loss of at most $\log L + C\eta$ for either component law. Thus the combined fine-label loss is at most $\log L + C\eta$, without an attention-smoothing error. The coarse head derivative is at least a fixed positive multiple of the combined coarse loss, by (77) and its relative loss estimates. Consequently

$$\dot{\mathcal{E}} \le -c(\mathcal{E} - p\log L - C\eta)_+^2, \qquad \mathcal{E}(s_0) \le p\log L + C\eta + \frac{C}{s_0 - t_*} = p\log L + o(p).$$

This establishes the same entry-loss bound (65).

For the upper incubation bound, $\chi \ge -o(1)$ and $b_s \ge cH^{-\gamma} e^{\chi}$. Choose a fixed sufficiently small $w_* > 0$. For $0 \le w_\nu \le w_*$ and $u_\nu \le B$, (78) gives

$$D_\nu \dot{w}_\nu \ge cpH^{-\gamma} - C\eta w_* \ge \tfrac{1}{2} cpH^{-\gamma} > 0.$$

The last step uses precisely $\eta \le pH^{-\gamma}$ and the choice of w_*. The head reaches w_* within CH^γ/p, unless it has already hit the prescribed margin, and cannot cross this head level downwards while these bounds hold. Thereafter (98) and (96) imply

$$\dot{\chi} \ge cpH^{-\gamma} e^{\chi}, \qquad \frac{\mathrm{d}}{\mathrm{d}t} e^{-\chi} \le -cpH^{-\gamma}.$$

This forces a fixed-margin hit within another CH^γ/p. Thus (63) holds with the full noise bound in the theorem; no extra small multiplicative constant in that noise bound is needed.

H.4 The transition and permanent success under noise

Choose B_0 and ε_* as in the noiseless proof, and let $\tau_{\nu,0}$ be the first clean-margin hit $u_\nu = B_0$ along the noisy trajectory. On the band $u_\nu \in [B_0/2, B_0]$, (60) and (83) give

$$\|\nabla u_\nu\| \le Cu_\nu, \qquad |\partial_{v_\nu} \mathcal{L}_s| \ge cpu_\nu.$$

Use the last crossing of $B_0/2$ before this first hit of B_0. The same fixed metric and reference-margin gradient give $-\dot{\mathcal{E}} \ge cp|\dot{u}_\nu|$ throughout that crossing interval. Integrating its total variation as in (68), and using the noisy entry-loss bound (65), yields, after reducing δ_0 if necessary,

$$\mathcal{E}(\tau_{\nu,0}) \le p(\log L - \delta_0).$$

The exact monotonicity of $\mathcal{E}$, together with (82), now implies $\ell_L(u_\nu) \le \log L - \delta_0/2$ on the continued bounded-margin interval, for sufficiently small p. Thus $u_\nu \ge b_- > 0$ there.

Differentiating the clean reference margin along the noisy flow gives the explicit counterpart of (71):

$$\begin{aligned}\dot{u}_\nu = \frac{p}{D_\nu} b_s^2 r_L(u_\nu)(1+o(1)) - \frac{L-1}{LD_\nu}\eta u_\nu(1+o(1)) \\ - u_\nu\left(\frac{a_s}{K} - \frac{c_s}{M}\right) G \\ + p u_\nu^2 r_L(u_\nu)\left(\frac{a_s^2}{K} + \frac{(1-b_s)^2}{L} + \frac{c_s^2}{M}\right) + o(p).\end{aligned} \tag{103}$$

The remainders are uniform on each fixed bounded margin interval. The damping is nonpositive; the last remainder comes from the query errors in (96). Dropping the two nonpositive terms and using $\sup_{u\geq 0} u^2 r_L(u) < \infty$ gives

$$\dot{u}_\nu \leq C_1 p,$$

with a leading constant independent of the fixed upper margin. On a fixed positive band, the head contribution has magnitude at most $C(p/\log H + \eta) = o(p)$, whereas the positive nuance-query term is of order p. As in the noiseless case, the opposing G term prevents this comparison from being a pointwise lower bound on the net margin velocity.

Choose B_1 with $\ell_L(B_1) \leq \varepsilon_*/4$. As long as $\mathcal{E} > \varepsilon_* p$, either $u_\nu \in [b_-, B_1]$, in which case (83) gives a query-gradient magnitude at least cp, or $u_\nu > B_1$, in which case (82) and (77) give a coarse-head gradient magnitude at least cp. Therefore $-\dot{\mathcal{E}} \geq c_1 p^2$ until the target sublevel is reached. At $\tau_{\nu,0}$, (82) also gives $\mathcal{E}(\tau_{\nu,0}) \geq p(\ell_L(B_0) - o(1)) > \varepsilon_* p$, so the target has not already been reached. The constants C_1, c_1 are fixed by these global speed bounds and the fixed thresholds b_-, B_1, ε_*. Choose $C_2 > \log L/c_1$, then $B_{\max} > B_0 + C_1 C_2 + 1$, and a fixed horizon multiplier containing $\tau_{\nu,0} + C_2/p$. Only after these choices take p small enough for the error estimates on this region. The lower-margin bound rules out a return to zero; the upper-speed bound rules out an exit through $B_{\max}$ within C_2/p. Otherwise the dissipation bound would remove more than $p \log L$ of excess loss before success. Thus the first target-sublevel time $\tau_{\nu,1}$ exists and $\tau_{\nu,1} - \tau_{\nu,0} \leq C_2/p$. At this time (82) implies $\ell_L(u_\nu) \leq \varepsilon_* + o(1) \leq 2\varepsilon_*$, whereas $\ell_L(B_0) \geq 4\varepsilon_*$. The bounded upper margin speed consequently gives $\tau_{\nu,1} - \tau_{\nu,0} \geq c/p$ as well.

The warm-up accuracy approximation and (84) prove the hallucination assertions on $[\tau_\varepsilon, \tau_{\nu,0}]$, with the same choice of a fixed τ_ε and sufficiently small p_ε. That choice also ensures signal-concentrated mean attention on the warm-up portion of this interval: the warm-up parameters are close to the reference trajectory, and the mean-attention error in (89) tends to zero. On $[t_*, \tau_{\nu,0}]$, the stronger bounds $a_g, a_s = 1 - o(1)$ and $b_s = o(1)$ follow from noisy continuation and carry over to their means. Finally (37) is exact for noisy data, not an asymptotic expansion. Once $\mathcal{E} \leq \varepsilon_* p$, all three eventual-accuracy bounds hold for every later time, even after the parameters leave the region where the preceding Gaussian estimates were used.

All constants depend only on K, L, M, T, ε, and all small-p estimates were uniform for $0 < \eta \leq pH^{-\gamma}$. Taking the smaller of the noiseless and noisy choices of p_ε, and adjusting the two time constants, proves Theorem 2 in its stated form. □

Mechanism of the three phases. The proof separates shared coarse learning from fine-label learning: both mixture components train the coarse head, while the positive fine-label force carries p. The signal-only trajectory and the initial incubation estimates show the additional attention factor $H^{-\gamma}$, which delays the first fixed fine-label margin to order H^γ/p. During the controlled plateau, signal attention is concentrated and the fine-label margin remains small. The transition is not a switch to nuance-dominated attention: on fixed positive margin bands, (62) shows that head growth compensates for a still-small nuance weight. The query contribution and loss-dissipation argument then give an order-$1/p$ transition. Because τ_ε is fixed, the ratio of plateau duration to transition duration is of order H^γ. These conclusions hold over the theorem's full noise range, with noisy attention interpreted through (89); after $\tau_{\nu,1}$, permanence follows from loss monotonicity, not from continued attention estimates.

I Proof of Theorem 3: the intermediate-noise regime

Strategy. Noise on the overwhelmingly common general sequences opposes growth of the fine-label head. We put an invariant upper bound on that head when η is a sufficiently large multiple of

p. This bounds fine-label improvement on specified sequences at every time. Separately, a fixed initial training interval makes the total loss close to $\log L$; loss monotonicity then preserves high general-sequence coarse accuracy forever. The theorem's hypotheses have $p > 0$ and $\eta \geq C_\varepsilon p$, so there is no noiseless case within this regime.

Fix the prescribed ε and put $w_* = 4\varepsilon$. First assume $p \leq 1/2$ and $\eta \leq 1$; both will be enforced by the final choice of p_ε. At $w_\nu = w_*$, condition on the attention of a general sequence. There is no fine signal. In the notation of (34), on the Gaussian event $\max_{1\leq j\leq L} |g_j| \leq 1$, every fine logit is in $[-w_*, w_*]$, because $\eta A_2 \leq 1$. Therefore every fine softmax coordinate is at least e^{-2w_*}/L, and

$$1 - \sum_{j=1}^{L} q_j^2 = \sum_{i\neq j} q_i q_j \geq \frac{L-1}{L} e^{-4w_*}.$$

The probability of that Gaussian event is a strictly positive constant depending only on L. Thus there is $h_* > 0$, depending only on L, ε, such that the conditional expectation of the preceding quantity is at least h_*, for every attention vector. Since $A_2 \geq 1/T$, (34) proves

$$\partial_{w_\nu} \mathbb{E}_g \Phi_L(w_*, 0, A_2) \geq \frac{\eta w_* h_*}{T}.$$

For a specified sequence, the same identity gives the lower bound $\partial_{w_\nu} \Phi_L \geq -b \geq -1$, since its variance term is nonnegative. The coarse loss does not depend on w_ν. Hence

$$\partial_{w_\nu} \mathcal{L}_s|_{w_\nu = w_*} \geq (1-p)\frac{\eta w_* h_*}{T} - p. \tag{104}$$

Choose

$$C_\varepsilon > \max\left\{1, \frac{4T}{w_* h_*}\right\}. \tag{105}$$

Whenever $\eta \geq C_\varepsilon p$, the right side of (104) is positive. Together with the nonnegative-head invariance in Lemma F.1, this proves

$$0 \leq w_\nu(t) \leq 4\varepsilon \qquad \text{for all } t \geq 0.$$

Equations (30)–(31) and (32) now imply exactly

$$Q_\nu - \frac{Q_{\mu,s}}{L} = \mathbb{E}_s\left[P_K(w_\mu, a, A_2)\left(P_L(w_\nu, b, A_2) - \frac{1}{L}\right)\right],$$

$$0 \leq Q_\nu - \frac{Q_{\mu,s}}{L} \leq \frac{w_\nu}{4} \leq \varepsilon.$$

This proves the stated joint-accuracy improvement bound at every time, independently of the state of the query. The baseline is $Q_{\mu,s}/L$, the joint accuracy obtained by retaining the same coarse predictions and choosing the fine label uniformly. It is not a claim that Q_ν is near $1/L$: Theorem 3 does not assert high specified-sequence coarse accuracy, so its conclusion alone is not the full hallucination condition of Definition 1.

It remains to prove the general-sequence coarse claim. On the $p = \eta = 0$ reference trajectory, $\mathcal{L}(t) = \log L + \ell_K(u_g^0(t))$ and the last term tends to zero. Choose a fixed $\tau_\varepsilon > 0$ for which it is at most $\varepsilon/8$. The reduced population loss and its finite-time flow depend continuously on (p, η) at $(0, 0)$. This follows either from the Gaussian representation or from the bounded-moment differentiability used in the warm-up proof, now on a fixed time interval. Consequently there is a fixed $\delta > 0$ such that

$$0 \leq p, \eta \leq \delta \quad \Longrightarrow \quad \mathcal{L}(\boldsymbol{W}(\tau_\varepsilon), \boldsymbol{v}(\tau_\varepsilon)) \leq \log L + \varepsilon/4.$$

After fixing C_ε in (105), choose

$$0 < p_\varepsilon \leq \min\left\{\frac{1}{2}, \delta, \frac{\delta}{C_\varepsilon}, \frac{1}{C_\varepsilon}, \frac{\varepsilon}{4\log L}\right\}.$$

The theorem's upper bound $\eta \leq C_\varepsilon p_\varepsilon$ now ensures both the required continuity condition and $\eta \leq 1$. At τ_ε,

$$\mathcal{E} \leq \varepsilon/4 + p\log L \leq \varepsilon/2.$$

For every later time, (36) and $p \leq 1/2$ therefore give

$$1 - Q_{\mu,g} \leq \mathbb{E}_g[-\log q_\mu(k)] \leq \frac{\mathcal{E}}{1-p} \leq \varepsilon.$$

This proves Theorem 3 with the stated dependencies of all constants. □