A BAYESIAN PERSPECTIVE ON THE ROLE OF EPISTEMIC UNCERTAINTY FOR DELAYED GENERALIZATION IN IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning to generalize from finite data is a central problem in machine learning. In-context learning provides a concrete solution to this problem while exhibiting a striking phenomenon known as delayed generalization: transformers can perfectly fit observed task-input combinations long before they can infer a held-out task from in-context examples. Although prior work has substantially advanced our understanding of this phenomenon, the evolution of uncertainty during this transition remains comparatively less explored. We study this question in a modular-arithmetic benchmark with latent linear tasks, using Improved Variational Online Newton (IVON) and a Last-Layer Laplace approximation to estimate posterior-predictive epistemic uncertainty (EU). Across different levels of task diversity and training-input coverage, both posterior approximations show lower epistemic uncertainty in configurations that generalize to held-out tasks. Furthermore, we show that EU follows a non-monotone pattern of rise-peak-decline during training, with its peak preceding the late increase in accuracy, in held-out tasks. Finally, we attempt to formalize this behavior with a simplified Bayesian linear model that reproduces the qualitative dynamics, showing that delayed generalization and non-monotone uncertainty can arise from slow spectral modes of the training-data covariance.

1 INTRODUCTION

In-context learning allows transformers to solve new tasks from examples provided in their input context, without updating their parameters (Brown et al., 2020). However, this ability can emerge only late in training: a model may fit observed task–input combinations long before it learns to infer a held-out task from context (He et al., 2024). This delayed transition is an instance of *grokking*, in which generalization emerges long after the training data have been fit (Power et al., 2022). Prior work explains ICL as implicit Bayesian inference or Bayesian model averaging, with the learned behavior depending on the pretraining task distribution (Xie et al., 2022; Raventos et al., 2023; Zhang et al., 2025). Other accounts show that transformers can learn simple function classes and implement linear estimators, algorithm selection, or gradient descent in context (Garg et al., 2022; Akyurek et al., 2023; Bai et al., 2023; Von Oswald et al., 2023; Xie et al., 2025). These accounts characterize the learned computation, but not how posterior predictive uncertainty evolves during the transition to task inference. In this work, we ask how uncertainty (especially epistemic uncertainty) changes while the model acquires task inference.

We address this question with Bayesian Deep Learning (BDL), which represents parameter uncertainty through a posterior distribution and marginalizes it to form the posterior predictive distribution (Wilson & Izmailov, 2020; Papamarkou et al., 2024; Agostinelli et al., 2026). We quantify EU as the mutual information between the prediction and the model parameters under the posterior, estimated as predictive entropy minus expected conditional entropy (Houlsby et al., 2011; Hüllermeier & Waegeman, 2021; Wimmer et al., 2023). Recent work develops scaling laws for how predictive uncertainty changes with dataset and model size (Rosso et al., 2025). Because exact Bayesian inference is intractable for transformers, we compare two complementary approximations. Improved Variational Online Newton (IVON) learns a diagonal Gaussian posterior over all weights during training (Shen et al., 2024), whereas Last-Layer Laplace fits a post-hoc Gaussian posterior over the

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.

output layer (Daxberger et al., 2021). To study this question, we use the latent modular-arithmetic ICL setup of He et al. (2024), in which a decoder-only transformer infers a latent linear task from in-context examples. This is a nontrivial setting that allows us to distinguish memorization, input generalization, task generalization, and their combination by varying the diversity of training tasks and the coverage of input pairs, allowing us to isolate well the learning dynamics, and the corresponding uncertainty trajectories.

Contributions. In summary, we make three contributions to understanding the relationship between delayed generalization and posterior predictive uncertainty in modular-arithmetic ICL. **(1) We identify an epistemic-uncertainty signature of grokking.** EU falls as the model learns the corresponding form of generalization, and both posterior approximations yield low EU when sufficient task diversity enables held-out-task inference. **(2) We characterize how this signature unfolds during training.** When the model learns to generalize jointly to held-out tasks and inputs, EU rises, peaks, and then collapses, with the peak occurring before the late increase in accuracy. **(3) We provide a simplified account of these dynamics.** Using a simple Bayesian linear model, we reproduce both delayed generalization and the rise–peak–collapse uncertainty trajectory, clarifying how the two phenomena can arise from the same learning process. A snapshot of these findings appears in Figure 1.

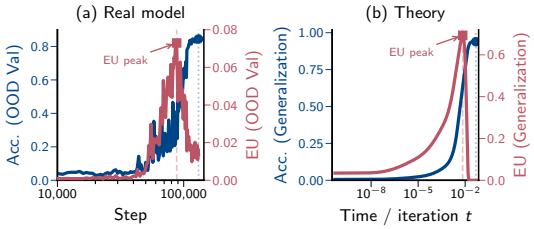


Figure 1: **A transition-aligned epistemic-uncertainty signature.** In the shown transformer run (left), epistemic uncertainty peaks before and collapses around the rise in OOD validation accuracy. The Bayesian linear model (right) exhibits the same non-monotone structure.

Related Work. Several lines of work model the transformer forward pass as implicit Bayesian inference, ridge regression, or gradient-based adaptation (Xie et al., 2022; Akyurek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023; Xie et al., 2025; Zhang et al., 2025). Other studies examine calibration and uncertainty under prompting (Zhang et al., 2024; Ling et al., 2024; Wang et al., 2025). We instead use BDL to trace posterior predictive uncertainty while task-inference ability is learned. Approximate marginalization makes this trajectory measurable in neural networks (e.g. Graves, 2011; Blundell et al., 2015; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Maddox et al., 2019; Ritter et al., 2018). Grokking work has characterized delayed generalization, learned representations, and progress measures (Power et al., 2022; Liu et al., 2022; Mohamadi et al., 2024; Clauw et al., 2024; Golechha, 2024). Closest to our setting, He et al. (2024) study latent linear tasks under in- and out-of-distribution evaluation. We retain their benchmark and analyze its Bayesian learning dynamics and uncertainty trajectories. Appendix B provides a detailed review.

2 Latent-Task Benchmark and Posterior-Predictive Uncertainty

We study delayed generalization in a modular-arithmetic benchmark where memorization, input generalization, and task generalization can be evaluated separately. In this section, we describe the setting.

2.1 Modular-Arithmetic ICL Benchmark

Tasks and sequences. Following He et al. (2024), each latent task (a, b) maps an input pair (x, y) to

$$z = ax + by \pmod p, \qquad p = 29, \tag{1}$$

where $(a, b) \in \mathcal{T} \equiv \{0, \dots, p-1\}^2$ and $(x, y) \in \mathcal{X} \equiv \{0, \dots, p-1\}^2$. The task vector is never revealed. A sequence contains $K-1$ completed examples and a query,

$$(x_1, y_1, z_1), \dots, (x_{K-1}, y_{K-1}, z_{K-1}), (x_K, y_K, \mathbf{?}).$$

Each instance is tokenized as three numeric tokens, with the query supplying only (x_K, y_K). For position i, context $\boldsymbol{c}_i$ comprises preceding triples and (x_i, y_i); the model predicts z_i. It is trained on all answer positions, but evaluated on the final answer z_K, which directly tests task inference from context. To create the datasets, we split the task set $\mathcal{T}$ and input set $\mathcal{X}$ into in-distribution (ID) and

out-of-distribution (OOD) subsets, $\mathcal{T}_{\text{id}}, \mathcal{T}_{\text{ood}}$ and $\mathcal{X}_{\text{train}}, \mathcal{X}_{\text{test}}$. We cross in- and out-of-distribution task sets $\mathcal{T}_{\text{id}}, \mathcal{T}_{\text{ood}}$ with training and held-out input sets $\mathcal{X}_{\text{train}}, \mathcal{X}_{\text{test}}$. Training uses only ID Train $(\mathcal{T}_{\text{id}}, \mathcal{X}_{\text{train}})$; ID Val isolates new inputs for seen tasks, OOD Train isolates new tasks on seen inputs, and OOD Val requires both (thus providing the most difficult held-out evaluation).

2.2 Posterior-Predictive Uncertainty and Evaluation

For fixed model parameters $\boldsymbol{\theta}$, the Transformer $f(\boldsymbol{\theta}, \boldsymbol{c})$ induces the categorical likelihood $p(z \mid \boldsymbol{c}, \boldsymbol{\theta})$ over the answer token z given context $\boldsymbol{c}$. Given training sequences $\mathcal{D}$, Bayesian prediction marginalizes these model parameters:

$$p(z \mid \boldsymbol{c}, \mathcal{D}) = \int p(z \mid \boldsymbol{c}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) \, d\boldsymbol{\theta}.$$

Because the exact posterior is intractable, we approximate $q(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta} \mid \mathcal{D})$ for the posterior and we estimate the predictive distribution by Monte Carlo.

Two complementary approximations. IVON (Shen et al., 2024) jointly learns the predictor and a diagonal-Gaussian posterior over all weights by variational inference, whereas Last-Layer Laplace (Daxberger et al., 2021) fits a post-hoc Gaussian to a MAP transformer's final layer, holding its feature extractor fixed. Both are approximations to the true posterior, albeit with different assumptions and computational trade-offs. All the details are provided in Appendix C.

Uncertainty decomposition. For samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)} \sim q(\boldsymbol{\theta})$, let $\bar{p}_S(z \mid \boldsymbol{c}) = S^{-1} \sum_{s=1}^{S} p(z \mid \boldsymbol{c}, \boldsymbol{\theta}^{(s)})$. Total uncertainty is $\mathbb{H}[\bar{p}_S]$, aleatoric uncertainty is the mean sampled entropy, and their difference is epistemic uncertainty (EU) (Wimmer et al., 2023; Rosso et al., 2025; Hüllermeier & Waegeman, 2021):

$$\mathrm{EU}(\boldsymbol{c}) = \mathbb{H}[\bar{p}_S(z \mid \boldsymbol{c})] - \frac{1}{S} \sum_{s=1}^{S} \mathbb{H}\Big[p(z \mid \boldsymbol{c}, \boldsymbol{\theta}^{(s)})\Big]. \tag{2}$$

Here $\mathbb{H}$ is categorical Shannon entropy. EU is label-free but requires multiple posterior-predictive passes and depends on the approximation.

Model and evaluation. We use a six-layer, width-512 decoder-only transformer with four attention heads, an MLP widening factor of four, tied embeddings, and Rotary Positional Embeddings (Su et al., 2024). Related task sets follow the parallelogram procedure of He et al. (2024), and tasks in a batch share an input sequence to limit interference. At evaluation, posterior-sampled predictions yield accuracy and the uncertainty decomposition above; full implementation details are in Appendices C and E.

3 From Delayed Generalization to an Uncertainty Signature

This section examines epistemic uncertainty (EU) during delayed task generalization. We first locate OOD-generalizing regimes across task diversity and training-input coverage. We next compare EU across this grid under IVON and Last-Layer Laplace, trace split-specific accuracy and uncertainty trajectories, and examine IVON as the number of in-context examples changes.

3.1 Where OOD Generalization Emerges

We first ask which training regimes support generalization to held-out tasks. We evaluate final accuracy over a grid that varies the number of in-distribution training tasks, $T \in \{32, 64, 128, 256, 512\}$, and the fraction of input pairs used for training, $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. This sweep separates the effects of task diversity and input-pair coverage. Figure 2 maps the accuracy across the (T, α) grid for the three evaluation splits. For both IVON and Last-Layer Laplace, high OOD Train accuracy occurs at $T \in \{256, 512\}$, indicating that sufficiently diverse training tasks can support adaptation to unseen task vectors. OOD Val is more selective: high accuracy appears only in these high-task-diversity regimes and varies with the training-input fraction. This pattern is consistent with the observations of He et al. (2024). As expected, the two approximations behave differently in terms

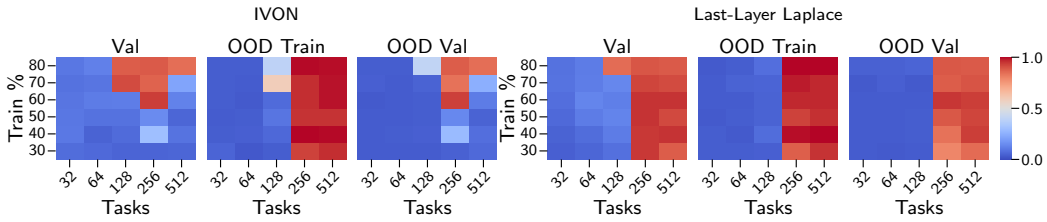


Figure 2: **Final accuracy across task-diversity regimes.** Both IVON and Last-Layer Laplace reach high OOD Train accuracy once task diversity is sufficiently large, indicating adaptation to unseen task vectors. High OOD Val accuracy appears only in a narrower high-diversity regime and depends on training-input coverage; this boundary is broader for Last-Layer Laplace than for IVON. Task diversity is therefore a shared requirement, while the posterior construction affects how far the generalizing region extends.

of the regions where high generalization accuracy is achieved. At the end of training, Last-Layer Laplace covers a broader high-accuracy region on OOD Val than IVON: its $T \in \{256, 512\}$ columns remain high across more α settings, whereas IVON's high OOD Val values are confined to selected fractions. Thus, the two posterior constructions agree that task diversity is important for OOD generalization, but they differ in the finer boundary of the high-accuracy region.

3.2 Epistemic Uncertainty Across Generalization Regimes

Having identified where OOD generalization emerges, we now ask how posterior-predictive uncertainty behaves across the same training regimes. We focus on task generalization: OOD Train evaluates held-out tasks on familiar inputs, whereas OOD Val tests whether this generalization persists when the inputs are also held out.

Figure 3 decomposes total predictive uncertainty into its epistemic (EU) and aleatoric (AU) components across the (T, α) grid for both IVON and Last-Layer Laplace. Because the two posterior approximations have different uncertainty scales, we compare their qualitative patterns rather than their absolute values. Despite their different posterior constructions, IVON and Last-Layer Laplace show the same qualitative pattern: mean EU is substantial at low task diversity and drops sharply at $T = 256$ and $T = 512$. This reduction coincides with the high-T regimes where the accuracy map first shows OOD generalization. Interestingly, across the (T, α) grid, the total predictive uncertainty (TU) does not show the same sharp drop as EU. Indeed, while EU drops at generalization, TU remains comparatively large, indicating that the reduction in EU is partially offset by an increase in AU. Matched configurations on opposite sides of the 0.75 OOD Val threshold make this distinction concrete (Table 1). For both posterior approximations, grokking is accompanied by a large reduction in EU but only a modest change in total uncertainty as higher AU offsets the lower EU. This pattern reinforces the idea that epistemic uncertainty is a more sensitive indicator of task generalization than total predictive uncertainty.

Table 1: **Matched-UQ contrasts.** Grokked configurations have lower EU despite comparable TU.

Method	Grokked?	(T, α)	TU	AU	EU
IVON	×	$(32, 0.8)$	1.527	0.700	0.827
	✓	$(256, 0.6)$	1.374	1.021 ↑ 46%	0.353 ↓ 57%
Laplace	×	$(32, 0.3)$	2.624	0.324	2.300
	✓	$(256, 0.3)$	2.255	1.524 ↑ 370%	0.731 ↓ 68%

3.3 Split-Specific Accuracy and Uncertainty Trajectories

The grid comparison locates low-EU at convergence, but does not determine how EU changes during training. Figure 4 provides two complementary views of IVON training trajectories: the left panel summarizes the ordering of accuracy and uncertainty events in runs that ultimately generalize on OOD Val, while the right panel compares how EU changes by the end of training across all runs. On the left, the characteristic sequence is that OOD Train reaches high accuracy first, OOD-Val EU then peaks, and OOD Val reaches high accuracy later. The peak therefore occurs during the transition to

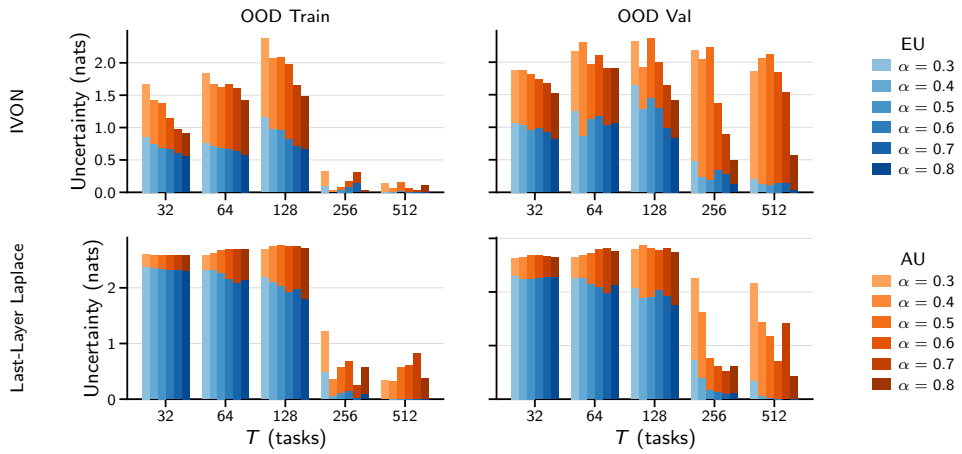


Figure 3: **OOD predictive-uncertainty decomposition across the same task-diversity grid.** At low task diversity, mean epistemic uncertainty (EU) is substantial on both OOD Train and OOD Val. Mean EU drops at larger T, qualitatively aligning with the region where the accuracy maps in Figure 2 first show high OOD accuracy, while the aleatoric component (AU) remains comparatively large. Bars report arithmetic means across available initialization and training seeds.

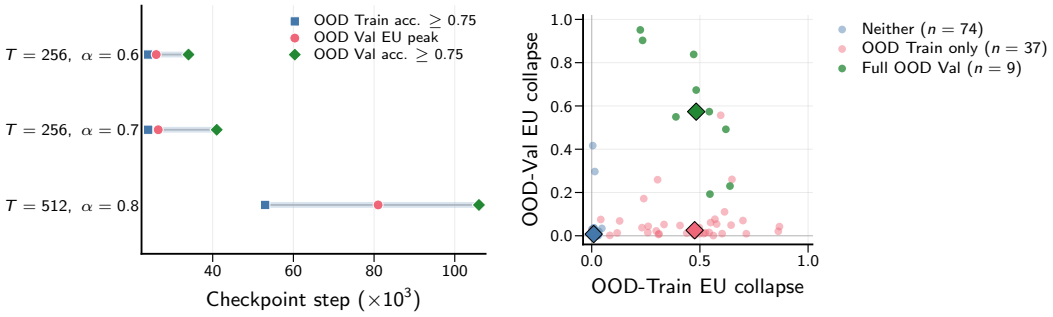


Figure 4: **All-run event timing and split-specific EU collapse.** *Left:* each row summarizes a (T, α) configuration that reaches high OOD-Val accuracy; its markers show the OOD-Train 0.75 crossing, OOD-Val EU peak, and OOD-Val 0.75 crossing. The typical ordering places the OOD-Val EU peak between the two accuracy transitions. *Right:* every point is one IVON run, colored by its endpoint generalization regime; diamonds are regime medians. Runs that generalize on both OOD splits show EU decline on both splits, while OOD-Train-only runs show this decline primarily on OOD Train.

the harder held-out-task-and-input generalization setting. This ordering is split-specific: OOD-Train EU does not show the same consistent peak-before-accuracy pattern.

On the right, the horizontal and vertical coordinates measure the EU decline on OOD Train and OOD Val, respectively. For each run and split, EU collapse is $1 - \mathrm{EU}_{\mathrm{end}}/\mathrm{EU}_{\mathrm{peak}}$, where $\mathrm{EU}_{\mathrm{peak}}$ is the maximum EU along the trajectory and $\mathrm{EU}_{\mathrm{end}}$ is EU at the end of training. The ratio $\mathrm{EU}_{\mathrm{end}}/\mathrm{EU}_{\mathrm{peak}}$ gives the endpoint EU as a share of the peak, so the collapse gives the share of peak EU lost by the endpoint. Values near 0 indicate little EU decline, while values near 1 indicate that EU has fallen close to zero. The panel shows little decline when neither OOD split generalizes, a decline concentrated on OOD Train when only that split generalizes, and a decline on both splits when OOD Val also generalizes. These event clocks are retrospective trajectory diagnostics and do not establish a prospective stopping rule. Threshold, smoothing, and total-/aleatoric-uncertainty controls are reported in Section A.1.

3.4 How In-Context Evidence Shapes Epistemic Uncertainty

We have seen that EU is a sensitive indicator of task generalization during training. Now we analyze the behavior of EU by varying test-time evidence, i.e., the number of in-context examples provided to the model before the query. Using IVON models trained at $(T = 256, \alpha = 0.8)$ with $K = 32$ triplets, we evaluate accuracy and EU on OOD Val as a function of the number of in-context examples, $K_{\text{test}} \in \{0, 1, \ldots, 31\}$ and we report the results in Figure 5. The figure maps OOD Val accuracy against EU collapse, with the reference peak computed over the context sweep for each seed. Median accuracy increases from 0.033 with zero in-context examples to 0.871 with all 31 examples.

The three groups follow the context path in sequence. With at most two in-context examples, the model has low OOD Val accuracy and low raw EU; these points can still have positive collapse because the normalization uses the later EU peak. The intermediate group contains the EU peak, where collapse is zero, before OOD Val accuracy reaches the high-accuracy region. After the accuracy threshold is crossed, many in-context examples are accompanied by high accuracy and low EU, producing the largest collapse values.

Together with the preceding training-time trajectories, this result is interesting because the OOD Val EU rise–peak–collapse pattern also emerges as task evidence accumulates at test time.

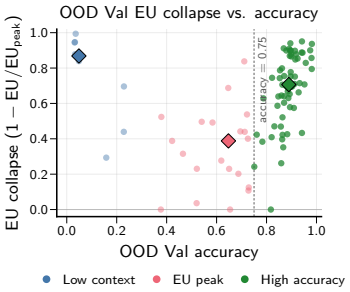


Figure 5: **Three context phases in the OOD Val accuracy–EU path.** The colors group the context sweep into low context ($K_{\text{test}} \leq 2$), an intermediate accuracy region, and a high-accuracy region (> 0.75). With low context, OOD Val accuracy and raw EU are both low. EU reaches its maximum in the intermediate region, so EU collapse is zero at the peak; this peak precedes the large increase in accuracy. At high accuracy, EU is low again and its collapse relative to the peak is large.

4 A Tractable Model for the Uncertainty Dynamics of Grokking

The empirical analysis in Section 3 leaves a specific mechanistic question: *when and why does uncertainty peak before grokking?* Making this relationship rigorous for the full transformer on modular arithmetic is currently out of reach. To isolate this relationship analytically, we follow Levi et al. (2024) and replace the sequence model with Bayesian linear regression. A linear regression model already captures delayed generalization and we will show that it provides tractable dynamics to reproduce three qualitative features from Section 3: empirical performance can improve before generalization performance, EU can rise and fall along the same learning trajectory, and its peak can precede a late-accuracy threshold.

This section develops the argument in four steps, with the overarching goal of connecting the empirical-generalization delay to the non-monotone epistemic uncertainty. First, we cast the variational inference dynamics as a system of ordinary differential equations (ODEs) and we derive closed-form solutions for the mean and covariance trajectories. Second, we compute the empirical and generalization risks along that trajectory, showing how the two differ in their behavior. Third, using Marchenko–Pastur asymptotics, weidentify the slow modes responsible for delayed generalization. Finally, by turning the model into a thresholded Bernoulli success event, we identify the non-monotone uncertainty curve and show that its peak can precede a late-accuracy threshold.

Appendix D provides the complete ODE derivation, finite-dimensional spectral sums, limiting calculation, and uncertainty proof.

Problem setup. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be a design matrix whose rows $\boldsymbol{x}_i^\top$ satisfy $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, and let a linear model with parameter $\boldsymbol{w}_\star$ generate the target vector $\boldsymbol{y} \in \mathbb{R}^n$. Consider the following model

$$p(\boldsymbol{y} \mid \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}_n), \qquad p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \tau^2 \boldsymbol{I}_d). \tag{3}$$

We study the regime $n, d \to \infty$ with fixed aspect ratio $\lambda = d/n < 1$. We approximate the posterior at time t by $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ and define

$$\boldsymbol{A} = \sigma^{-2} \boldsymbol{X}^\top \boldsymbol{X} + \tau^{-2} \boldsymbol{I}_d, \qquad \boldsymbol{b} = \sigma^{-2} \boldsymbol{X}^\top \boldsymbol{y}. \tag{4}$$

Here $\boldsymbol{A}$ is the posterior precision matrix, and $\boldsymbol{b}$ is the precision-weighted data term. The matrix $\boldsymbol{A}$ also governs the rate at which the variational parameters approach their stationary values. For this

model and Gaussian variational family, the Bures–Wasserstein gradient flow of Lambert et al. (2022) yields the closed ODEs

$$\dot{\boldsymbol{\mu}}_t = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{\mu}_t, \qquad \dot{\boldsymbol{\Sigma}}_t = 2\boldsymbol{I}_d - \boldsymbol{A}\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t\boldsymbol{A}. \tag{5}$$

Let $\boldsymbol{\mu}_\infty = \boldsymbol{A}^{-1}\boldsymbol{b}$ and $\boldsymbol{\Sigma}_\infty = \boldsymbol{A}^{-1}$ denote the stationary mean and covariance. Defining $\boldsymbol{E}_t = e^{-\boldsymbol{A}t}$, $\delta\boldsymbol{\mu}_0 = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_\infty$, and $\Delta\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_\infty$, the ODEs have solutions

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_\infty + \boldsymbol{E}_t\delta\boldsymbol{\mu}_0, \qquad \boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_\infty + \boldsymbol{E}_t\Delta\boldsymbol{\Sigma}_0\boldsymbol{E}_t.$$

Along this trajectory, we track the expected empirical and generalization risks

$$\mathcal{L}_{\text{emp}}(t) = \mathbb{E}_{q_t}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, \boldsymbol{x}_i^\top\boldsymbol{w})\right], \qquad \mathcal{L}_{\text{gen}}(t) = \mathbb{E}_{q_t(\boldsymbol{w})p(\boldsymbol{x},y)}\left[\ell(y, \boldsymbol{x}^\top\boldsymbol{w})\right], \tag{6}$$

where $\ell(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2$. For a Gaussian predictor, evaluating the quadratic loss gives

$$\mathcal{L}_{\text{emp}}(t) = \frac{1}{2n}\left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_t\|^2 + \operatorname{Tr}\Big(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Sigma}_t\Big)\right)$$
$$\mathcal{L}_{\text{gen}}(t) = \frac{1}{2}\left(\|\boldsymbol{\mu}_t - \boldsymbol{w}_\star\|^2 + \operatorname{Tr}(\boldsymbol{\Sigma}_t)\right).$$

For $r \in \{\text{emp}, \text{gen}\}$, the excess risk is the difference between the risk at time t and its limiting value at the stationary variational solution: $\widetilde{\mathcal{L}}_r(t) = \mathcal{L}_r(t) - \mathcal{L}_r(\infty)$. It isolates the time-dependent component of the risk after subtracting the stationary contribution and converges to zero as $t \to \infty$.

Risk dynamics and spectral decomposition. Substituting the mean and covariance trajectories and subtracting their stationary values separates each excess risk into a term from the mean gap and a term from the covariance gap. In the matched-interpolation limit, $\boldsymbol{\mu}_\infty \to \boldsymbol{w}_\star$ and the normalized stationary training residual vanishes as $n, d \to \infty$ with $d/n = \lambda < 1$. The terms linear in $\boldsymbol{E}_t$ are therefore $o_n(1)$, leaving

$$\widetilde{\mathcal{L}}_{\text{emp}}(t) = \frac{1}{2n}\delta\boldsymbol{\mu}_0^\top\boldsymbol{E}_t\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{E}_t\delta\boldsymbol{\mu}_0 + \frac{1}{2n}\operatorname{Tr}\Big(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{E}_t\Delta\boldsymbol{\Sigma}_0\boldsymbol{E}_t\Big) + o_n(1), \tag{7}$$

$$\widetilde{\mathcal{L}}_{\text{gen}}(t) = \frac{1}{2}\delta\boldsymbol{\mu}_0^\top\boldsymbol{E}_t^2\delta\boldsymbol{\mu}_0 + \frac{1}{2}\operatorname{Tr}(\boldsymbol{E}_t\Delta\boldsymbol{\Sigma}_0\boldsymbol{E}_t) + o_n(1). \tag{8}$$

In each expression, the first term measures the residual error in the variational mean and the second measures the contribution of the covariance gap. Both risks evolve under the same operator $\boldsymbol{A}$, but the empirical risk inserts the training covariance $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^\top\boldsymbol{X}/n$, whereas the generalization risk inserts the identity.

To make the spectral difference explicit, diagonalize $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{U}\operatorname{diag}(\nu_1,\ldots,\nu_d)\boldsymbol{U}^\top$, where $\nu_1,\ldots,\nu_d$ are the eigenvalues of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}$. The operator $\boldsymbol{A}$ has the same eigenvectors and eigenvalues $a_i = (n/\sigma^2)\nu_i + 1/\tau^2$. Consequently, empirical excess risk contains mode-wise terms proportional to $\nu_i e^{-2a_i t}$, whereas generalization excess risk contains the corresponding $e^{-2a_i t}$ terms without the factor ν_i. Under the self-averaging initialization assumption stated in Appendix D, the mode-dependent initial weights can be replaced by their average. Define

$$C_0 \equiv \frac{1}{2}\left(\|\delta\boldsymbol{\mu}_0\|^2 + \operatorname{Tr}(\Delta\boldsymbol{\Sigma}_0)\right); \tag{9}$$

Under this assumption, the combined mean and covariance contribution of each eigenmode has asymptotic weight C_0/d. Consequently, Equations (7) and (8) reduce at leading order to the finite spectral averages

$$\widetilde{\mathcal{L}}_{\text{emp}}(t) \approx \frac{C_0}{d}\sum_{i=1}^{d}\nu_i\exp\left[-2\left(\frac{n}{\sigma^2}\nu_i + \frac{1}{\tau^2}\right)t\right], \quad \widetilde{\mathcal{L}}_{\text{gen}}(t) \approx \frac{C_0}{d}\sum_{i=1}^{d}\exp\left[-2\left(\frac{n}{\sigma^2}\nu_i + \frac{1}{\tau^2}\right)t\right].$$

For the Gaussian configuration we have, the empirical spectral measure $d^{-1}\sum_{i=1}^{d}\delta_{\nu_i}$ converges to the Marchenko–Pastur law $\text{MP}(\lambda)$. Writing either summand as $f_r(\nu_i)$ therefore gives

$$\frac{1}{d}\sum_{i=1}^{d}f_r(\nu_i) \longrightarrow \mathbb{E}_{\nu\sim\text{MP}(\lambda)}[f_r(\nu)],$$

which gives the limiting expressions

$$\widetilde{\mathcal{L}}_{\text{emp}}(t) \approx C_0\, \mathbb{E}_{\nu\sim \text{MP}(\lambda)} \left[\nu \exp\left(-2 \left(\frac{n}{\sigma^2}\nu + \frac{1}{\tau^2} \right) t \right) \right], \tag{10}$$

$$\widetilde{\mathcal{L}}_{\text{gen}}(t) \approx C_0\, \mathbb{E}_{\nu\sim \text{MP}(\lambda)} \left[\exp\left(-2 \left(\frac{n}{\sigma^2}\nu + \frac{1}{\tau^2} \right) t \right) \right]. \tag{11}$$

The extra factor ν suppresses modes near the lower spectral edge in empirical risk. Those modes are simultaneously the slowest to decay, so training can appear complete while they continue to dominate generalization.

The lower Marchenko–Pastur edge $\nu_- = (1 - \sqrt{\lambda})^2$ controls the late-time behavior:

$$\widetilde{\mathcal{L}}_r(t) \sim K_r t^{-3/2} e^{-\kappa t}, \qquad r \in \{\text{emp}, \text{gen}\}, \qquad \kappa = 2\left(\frac{n}{\sigma^2}(1 - \sqrt{\lambda})^2 + \frac{1}{\tau^2} \right), \tag{12}$$

with prefactor ratio $K_{\text{gen}}/K_{\text{emp}} = (1 - \sqrt{\lambda})^{-2}$. The two risks share the exponential rate κ, but the empirical weighting changes the prefactor. This mismatch produces the empirical–generalization delay used below.

Threshold-model uncertainty and delayed EU peaks. The spectral analysis above explains how slow modes delay the decay of empirical and generalization risk. We now ask whether these risk trajectories can also generate the rise–peak–collapse pattern in EU and place its maximum before high accuracy. Following Levi et al. (2024), we connect the continuous regression residual to accuracy by declaring a prediction successful whenever its squared residual is at most a tolerance $\epsilon > 0$. We then use posterior ambiguity about this success event to define EU. Accuracy and EU can consequently be expressed in terms of the same excess-risk trajectory, making the timing of the uncertainty peak and the late-accuracy crossing analytically comparable. For a point $(\boldsymbol{x}, y)$, define the thresholded success variable

$$Z_\epsilon(\boldsymbol{x}, y; \boldsymbol{w}) = \mathbb{1}\{(y - \boldsymbol{x}^\top \boldsymbol{w})^2 \le \epsilon\}. \tag{13}$$

Taking its expectation under the variational and relevant data distributions defines empirical and generalization accuracies $\mathcal{A}_{\text{emp}}$ and $\mathcal{A}_{\text{gen}}$; Appendix D gives the formal definitions. Define t_{grok} as the first time that accuracy reaches 95%. Under the high-dimensional approximation, this is equivalent to $\widetilde{\mathcal{L}}(t_{\text{grok}}) \approx \epsilon/8$.

It remains to determine how uncertainty about the thresholded event evolves as excess risk decreases. The following lemma derives this relationship and identifies the risk level at which uncertainty is maximal.

Lemma 4.1 (Epistemic uncertainty of the threshold model)**.** *For fixed $(\boldsymbol{x}, y)$, the residual under q_t is Gaussian and the thresholded success event is Bernoulli. Conditional on $\boldsymbol{w}$, the event is deterministic, so its conditional entropy vanishes and its epistemic uncertainty equals the binary entropy of its posterior success probability. In the high-dimensional limit,*

$$\text{EU}(t) \approx h_2\left(\text{erf}\left(\sqrt{\frac{\epsilon}{4\widetilde{\mathcal{L}}(t)}} \right) \right), \qquad h_2(p) = -p \log p - (1-p)\log(1-p). \tag{14}$$

This entropy is non-monotone in the decreasing excess risk and peaks when

$$\widetilde{\mathcal{L}}(t_{\text{peak}}) = \frac{\epsilon}{4(\text{erf}^{-1}(1/2))^2} \approx 1.10\,\epsilon, \tag{15}$$

where the posterior success probability is $1/2$ and the thresholded event is maximally ambiguous.

Remark 4.1. Because excess risk decreases over time and $1.10\,\epsilon > \epsilon/8$, the threshold-model EU peak precedes the 95% accuracy crossing. Similar results can be derived for other accuracy thresholds.

Finally, we analyze the delay between the empirical and generalization EU peaks as an uncertainty-based counterpart of the accuracy-defined grokking delay.

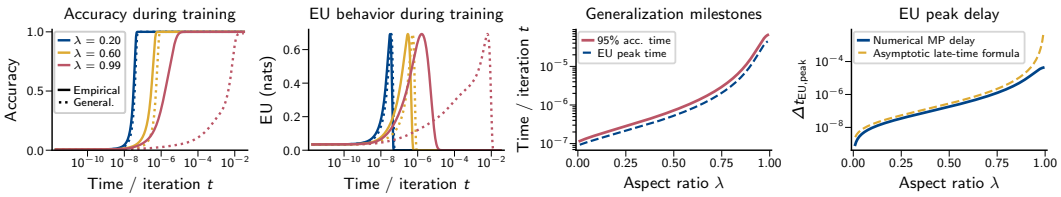


Figure 6: **Training and generalization dynamics of the simplified model.** Increasing the aspect ratio λ delays generalization relative to training (*left*). Threshold-model EU rises and falls along the same trajectory (*center left*), and its maximum precedes the 95% accuracy crossing (*center right*). The empirical–generalization EU-peak separation grows as $\lambda \to 1$ and is approximated by the lower-edge asymptotic in Theorem 4.1 (*right*).

Theorem 4.1 (Delay between threshold-uncertainty peaks)**.** *Assume the construction of Lemma 4.1 and the spectral asymptotics above. If the empirical and generalization uncertainty curves peak in this regime, then their separation satisfies*

$$\Delta t_{\mathrm{EU,peak}}(\lambda) = t_{\mathrm{peak,gen}} - t_{\mathrm{peak,emp}} \simeq \frac{\log\left(\frac{1}{1-\sqrt{\lambda}}\right)}{\frac{n}{\sigma^2}(1-\sqrt{\lambda})^2 + \frac{1}{\tau^2}}. \tag{16}$$

The theorem turns the qualitative connection into a scaling prediction: the separation between the empirical and generalization EU peaks increases as the lower spectral edge approaches zero. Within the simplified model, this connects uncertainty timing to the geometry of the training data, since the same slow covariance modes delay generalization and the generalization EU peak. The visual example below illustrates this shared spectral origin across the four stages of the analysis.

Visual summary of the simplified model. Figure 6 assembles the preceding results into a single left-to-right analysis. Its four panels follow the construction from the spectral excess-risk dynamics to the resulting accuracy and uncertainty event times. The left panel evaluates the empirical and generalization spectral averages as λ increases, showing a widening delay between their accuracy curves. The center-left panel maps the corresponding risk trajectories through the entropy of the Bernoulli distribution, revealing the rise and subsequent collapse of empirical and generalization EU. The center-right panel compares the two event thresholds: EU reaches its maximum at $\widetilde{\mathcal{L}} \approx 1.10\,\epsilon$, before the risk falls to $\epsilon/8$ at 95% accuracy, meaning that the EU peak precedes the late-accuracy crossing at various λ. The right panel varies λ and compares the resulting empirical–generalization EU-peak separation with the lower-edge approximation in Theorem 4.1, showing good agreement between the two until $\lambda \to 1$. Together, the panels show how slow spectral modes produce delayed generalization and an earlier uncertainty peak within the simplified model.

5 Conclusion and Limitations

We studied how epistemic uncertainty evolves during delayed generalization in in-context learning. The results reveal a transition-aligned signature in which uncertainty rises, peaks before late generalization, and declines as task generalization emerges. This signature connects posterior uncertainty with changes in how the model uses in-context evidence. A tractable Bayesian model shows how slow learning directions can jointly produce delayed generalization and non-monotone uncertainty. These findings suggest that epistemic uncertainty offers a useful perspective on the learning dynamics of grokking.

Limitations. Whether the same uncertainty dynamics arise in natural language, large pretrained models, and other forms of in-context learning remains unknown. The causal relationship between epistemic uncertainty and delayed generalization is unresolved, including whether uncertainty drives, constrains, or records the transition. A complete theory must connect the spectral account in the simplified model to representation learning and posterior geometry in transformers.

AI USE STATEMENT

We used generative AI tools for software-engineering and editorial assistance, specifically code implementation support and review beyond simple autocompletion, prose revision and manuscript organization. The tools did not originate research ideas or scientific claims. Math derivations and proofs were developed by the authors, and double-checked for correctness using various LLMs and AI agents. We reviewed all AI-assisted work and take responsibility for the final study, text, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

We document task construction, batching, and posterior inference in Appendix C, and collect the numerical architecture, optimization, evaluation, and experiment-grid settings in Appendix E; Appendix D gives the full derivations and proofs for the simplified model. Model checkpoints and code for reproducing the experiments will be released at a later date.

REFERENCES

Claudio Agostinelli, Laurence Aitchison, Emanuel Aldea, Richard Allmendinger, Sebastian Ament, Ben Anson, Julyan Arbel, Eytan Bakshy, Max Balandat, Nicola Bariletto, Freddie Bickford Smith, Battista Biggio, Michele Caprio, Neil Chada, Wenlin Chen, Wenlong Chen, Nathaël Da Costa, Nico Daheim, Andreas Damianou, Michael Deistler, David Dunson, Nikita Durasov, Maurizio Filippone, Lucia Filippozzi, Gergely Flamich, Vincent Fortuin, Gianni Franchi, Giorgio Fumera, Yarin Gal, Wenbo Gong, Yuqi Gu, Cheng Guo, Nick Hauptvogel, Jiajun He, Philipp Hennig, José Miguel Hernández-Lobato, Nhat Ho, Clara Hoffmann, Aliaksandr Hubin, Christian Igel, Ajay Jasra, Matt Jones, Nadja Klein, Olivier Laurent, Kody J. H. Law, Emanuele Ledda, Bolian Li, Yingzhen Li, Xinzhu Liang, Jakob H Macke, Stephan Mandt, Luckeciano Carvalho Melo, Benjamin Kurt Miller, Edward Milsom, Bruno Mlodozeniec, Bálint Mucsányi, Samuel Müller, Kevin Murphy, Eric Nalisnick, Marcos Negre Saura, Christopher Nemeth, Huy Nguyen, Khai Nguyen, Shreyas Padhy, Konstantina Palla, Wei Pan, Theodore Papamarkou, Dhruvesh Patel, Alina Peluso, Simon Pepin Lehalleur, Konstantinos Pitas, Patrick Pynadath, Tom Rainforth, Maxime Robeyns, Fabio Roli, Breeshey Roskams-Hieter, Simone Rossi, Tim G. J. Rudner, David Rügamer, Dino Sejdinovic, Yevgeny Seldin, Torben Sell, Louis Sharrock, Sumeetpal Singh, Joanna Sliwa, Emanuel Sommer, Qifan Song, Ba-Hien Tran, Richard E. Turner, Tycho F. A. van der Ouderaa, Mark van der Wilk, Mariia Vladimirova, Sara Wade, Xi Wang, Tobias Weber, Susan Wei, Christopher K. Wikle, Andrew Gordon Wilson, Lisa Wimmer, Adam Yang, Yibo Yang, Myungsoo Yoo, Cheng Zhang, Likun Zhang, Ruqi Zhang, and Yuren Zhou. *Handbook of Bayesian Deep Learning*. Zenodo, August 2026. doi: 10.5281/zenodo.22086450. URL https://doi.org/10.5281/zenodo.22086450.

Ekin Akyurek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=liMSqUuVg9.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. In *International Conference on Machine Learning (ICML)*, May 2015. URL http://arxiv.org/abs/1505.05424.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Kenzo Clauw, Sebastiano Stramaglia, and Daniele Marinazzo. Information-Theoretic Progress Measures reveal Grokking is an Emergent Phase Transition. In *ICML 2024 Workshop on Mutual Information (MI)*, 2024. doi: 10.48550/arXiv.2408.08944.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux - Effortless Bayesian Deep Learning. In *neurips*, volume 34, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, 2016.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Satvik Golechha. Progress Measures for Grokking on Real-world Tasks. In *ICML 2024 Workshop on High-dimensional Learning Dynamics (HiLD)*, 2024. doi: 10.48550/arXiv.2405.12755.

Alex Graves. Practical Variational Inference for Neural Networks. In *Advances in Neural Information Processing Systems*, 2011.

Tianyu He, Darshil Doshi, Aritra Das, and Andrey Gromov. Learning to grok: Emergence of in-context learning and skill composition in modular arithmetic tasks, 2024. URL https://arxiv.org/abs/2406.02550.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning, 2011. URL https://arxiv.org/abs/1112.5745.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 3(110):457–506, 2021.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam. In *International Conference on Machine Learning (ICML)*, 2018. URL http://arxiv.org/abs/1806.04854.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems 30*, 2017. URL https://papers.nips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=K2PTuvVTF1L.

Noam Itzhak Levi, Alon Beck, and Yohai Bar-Sinai. Grokking in Linear Estimators – A Solvable Model that Groks without Understanding. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GH2LYb9XV0.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. Uncertainty Quantification for In-Context Learning of Large Language Models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3357–3370, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.184. URL https://aclanthology.org/2024.naacl-long.184/.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Processing Systems (NeurIPS)*, volume 32, 2019.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why Do You Grok? A Theoretical Analysis of Grokking Modular Addition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35155–35173. PMLR, 2024.

Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, José Miguel Hernández-Lobato, Aliaksandr Hubin, Alexander Immer, Theofanis Karaletsos, Mohammad Emtiyaz Khan, Agustinus Kristiadi, Yingzhen Li, Stephan Mandt, Christopher Nemeth, Michael A Osborne, Tim G. J. Rudner, David Rügamer, Yee Whye Teh, Max Welling, Andrew Gordon Wilson, and Ruqi Zhang. Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI. In *International Conference on Machine Learning*, 2024.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

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

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable Laplace approximation for neural networks. In *International conference on learning representations*, 2018.

Mattia Rosso, Simone Rossi, Giulio Franzese, Markus Heinonen, and Maurizio Filippone. Scaling Laws for Uncertainty in Deep Learning. *arXiv preprint arXiv:2506.09648*, 2025.

Yuesong Shen, Nico Daheim, Bai Cong, Peter Nickl, Gian Maria Marconi, Bazan Clement Emile Marcel Raoul, Rio Yokota, Iryna Gurevych, Daniel Cremers, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Variational Learning is Effective for Large Deep Networks. 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. 44665–44686. PMLR, 21–27 Jul 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Yifei Wang, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. Uncertainty Unveiled: Can Exposure to More In-context Examples Mitigate Uncertainty for Large Language Models? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20659–20678, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1062. URL https://aclanthology.org/2025.findings-acl.1062/.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *NeurIPS*, 2020.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Conference on Uncertainty in Artificial Intelligence*, 2023.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, 2022.

Shifeng Xie, Rui Yuan, Simone Rossi, and Thomas Hannagan. The Initialization Determines Whether In-Context Learning Is Gradient Descent. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=fvqSKLDtJi.

Hanlin Zhang, YiFan Zhang, Yaodong Yu, Dhruv Madeka, Dean Foster, Eric Xing, Himabindu Lakkaraju, and Sham Kakade. A Study on the Calibration of In-context Learning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6118–6136, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.340. URL https://aclanthology.org/2024.naacl-long.340/.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and How Does In-Context Learning Learn? Bayesian Model Averaging, Parameterization, and Generalization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

A Additional Experiments

For legibility, Figure 7 reproduces the task-diversity endpoint maps from Figure 2 at full width.

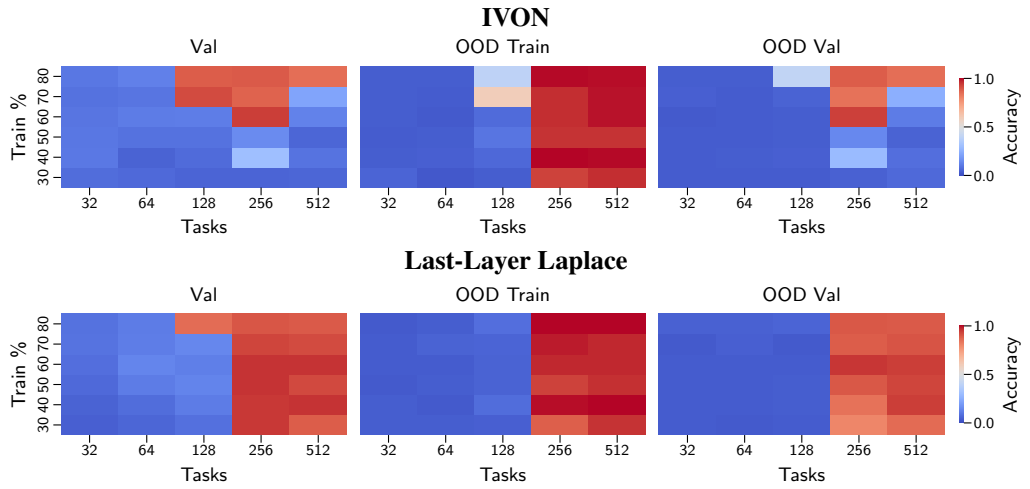


Figure 7: **Full-size task-diversity endpoint maps.** Final-checkpoint accuracy across effective task count T, training-input fraction α, and evaluation split. For each method, at least one seed attains high OOD Train accuracy at large T; high OOD Val accuracy is more selective and differs quantitatively between posterior constructions. This is the full-size version of Figure 2.

A.1 Trajectory-Analysis Controls

The main trajectory analysis uses a centered three-checkpoint median, a final-five endpoint median, and a retrospective 0.75 accuracy threshold. We vary the smoothing window and threshold below, always retaining the same nine runs selected by stable OOD-Val success under the main 0.75 endpoint criterion.

Table 2: **EU-collapse sensitivity to smoothing.** Entries are median OOD-Train/OOD-Val collapse percentages, with regime counts fixed by the main endpoint definition.

Endpoint regime	n	**No smoothing**	**3 checkpoints**	**5 checkpoints**
Neither	74	1.8/1.5	0.9/0.8	0.4/0.4
OOD Train only	37	52.0/7.2	47.6/2.6	39.1/2.2
Full OOD Val	9	56.6/68.3	48.3/57.4	44.2/55.9

Table 3: **OOD-Val peak-order sensitivity among the nine stable successes.** Each entry gives the number whose EU peak precedes the stated accuracy crossing, followed in parentheses by the median crossing-minus-peak time.

Smoothing window	0.50 **crossing**	0.75 **crossing**	0.80 **crossing**
1 checkpoint	5/9 (1k)	7/9 (8k)	7/9 (12k)
3 checkpoints	6/9 (2k)	8/9 (10k)	8/9 (13k)
5 checkpoints	7/9 (3k)	8/9 (11k)	8/9 (14k)

Table 4: **EU versus total- and aleatoric-uncertainty peak locations.** Entries are median peak positions as percentages of each run's evaluation trajectory, reported as EU/TU/AU for the centered three-checkpoint median.

Endpoint regime	n	**OOD Train EU/TU/AU**	**OOD Val EU/TU/AU**
Neither	74	$100/1/2$	$99/1/2$
OOD Train only	37	$44/3/3$	$97/3/3$
Full OOD Val	9	$40/3/3$	$34/3/3$

The collapse distinction between regimes persists under all three smoothing choices, although its magnitude changes. Peak-before-crossing counts vary with both the late-accuracy threshold and smoothing, which is why the main text reports the exact 0.75/three-checkpoint convention rather than treating the event order as universal. Finally, total and aleatoric uncertainty generally peak near the start of training, whereas EU peaks later and in a regime- and split-dependent location; the reported EU trajectory is therefore not a restatement of total predictive entropy.

A.2 In-Context Example Count and y_i-Token Corruption

At fixed model weights, we vary the number of completed in-context examples and, separately, corrupt their y_i tokens. These evaluations test whether EU responds to the examples from which the model infers the latent task.

Number of completed in-context examples. Models are trained with $K = 32$ triplets—31 completed in-context examples and one query—and evaluated at fixed weights with $K \in \{1, \ldots, 32\}$. In both posterior approximations, OOD Val accuracy generally rises with $K - 1$, while EU first rises and then falls, on method-dependent scales. The center of Figure 8 plots the same IVON sweep in the accuracy–EU plane; its path parameter is the number of completed examples, not checkpoint step.

Corruption of y_i tokens in completed in-context examples. At fixed weights, we replace each completed example's y_i token with a uniformly sampled token with probability $p_{\text{noise}} \in \{0, 0.1, \ldots, 1\}$; the x_i and z_i tokens, final query, and evaluation targets remain clean. In the right block of Figure 8, mean accuracy falls and mean EU rises under both posterior approximations, with a larger EU fold-change under Laplace.

Result. At fixed weights, EU responds to the number of completed in-context examples and corruption of their y_i tokens; this result does not identify a cause of the training-time transition.

Figures 9 and 11 provide the full split-wise accuracy, EU, and AU results underlying the compact OOD Val summary in Figure 8.

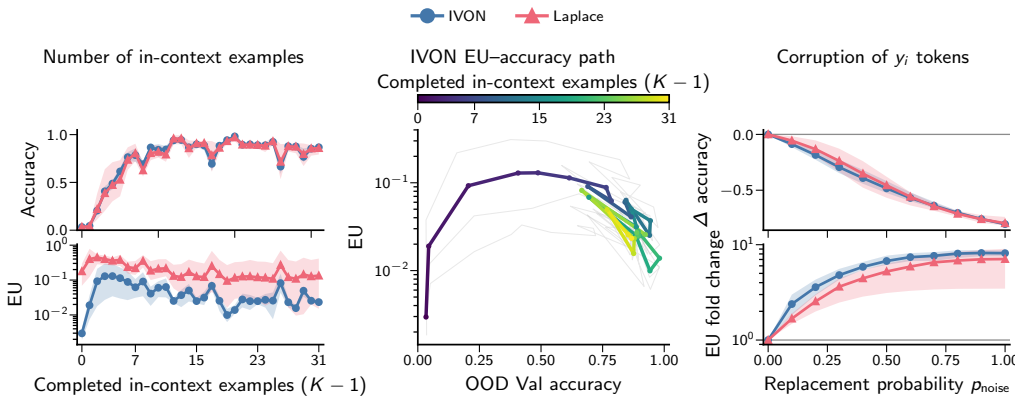


Figure 8: **Accuracy and epistemic uncertainty when varying the number of completed in-context examples or corrupting their y_i tokens.** Fixed models trained at $(T = 256, \alpha = 0.8, K = 32)$ are evaluated with 64 posterior samples. *Left:* OOD Val accuracy and EU versus the number $K - 1$ of completed examples. *Center:* the same IVON sweep in the accuracy–EU plane; faint lines are seeds, the colored path is their mean, and color encodes $K - 1$. *Right:* within-seed accuracy change and EU fold-change as each completed example's y_i token is replaced with probability p_{noise}, relative to $p_{\text{noise}} = 0$. Other curves are seed means; bands are empirical 2.5th–97.5th percentile ranges, not confidence intervals (IVON seeds 0, 1, 3; Laplace seeds 0–3). Figures 9 and 11 give full split-wise results.

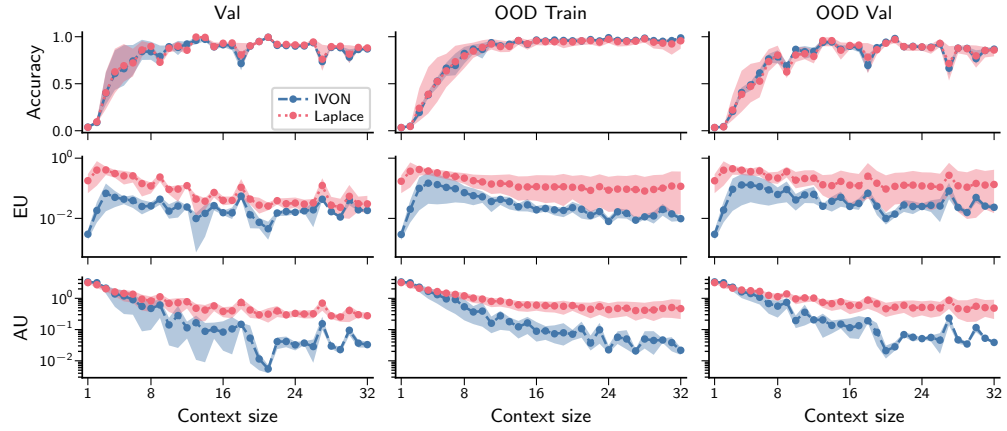
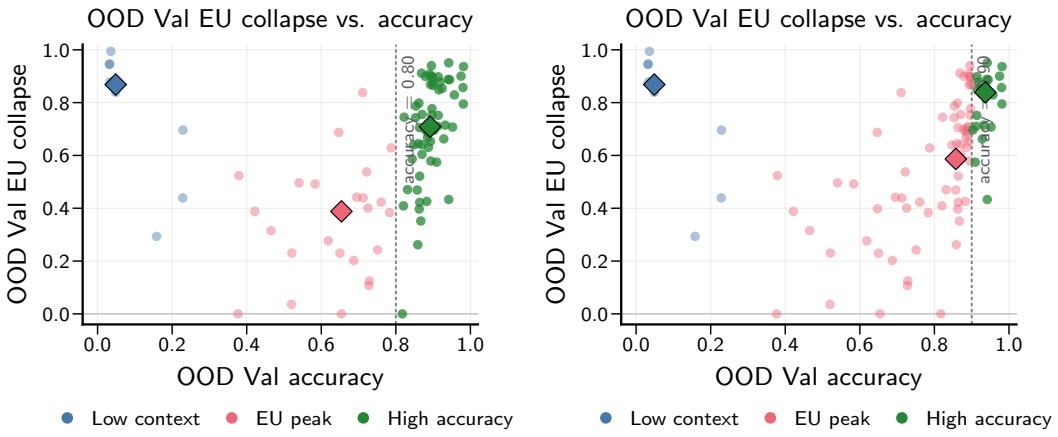


Figure 9: **Full split-wise evaluation across numbers of completed in-context examples.** Accuracy, EU, and AU as the number $K - 1$ of completed in-context examples increases from zero to 31 for fixed models trained at $(T = 256, \alpha = 0.8, K = 32)$. Mean accuracy rises and mean AU falls, while mean EU follows a non-monotone trajectory. Curves are arithmetic seed means, and bands span empirical 2.5th–97.5th percentiles across seeds, not confidence intervals; IVON uses seeds 0, 1, and 3, Laplace uses seeds 0–3, and both use 64 posterior samples.

For threshold sensitivity, Figure 10 repeats the OOD Val EU-collapse view with accuracy guides at 0.8 and 0.9.



(a) Accuracy guide at 0.8.

(b) Accuracy guide at 0.9.

Figure 10: **Threshold comparison for OOD Val EU collapse.** The three-phase pattern persists as the accuracy guide becomes stricter. The guide is crossed by $62/96$ context evaluations at 0.8 and $20/96$ at 0.9. The EU-peak region remains closest to zero collapse, while the high-accuracy region contains the larger collapse values.

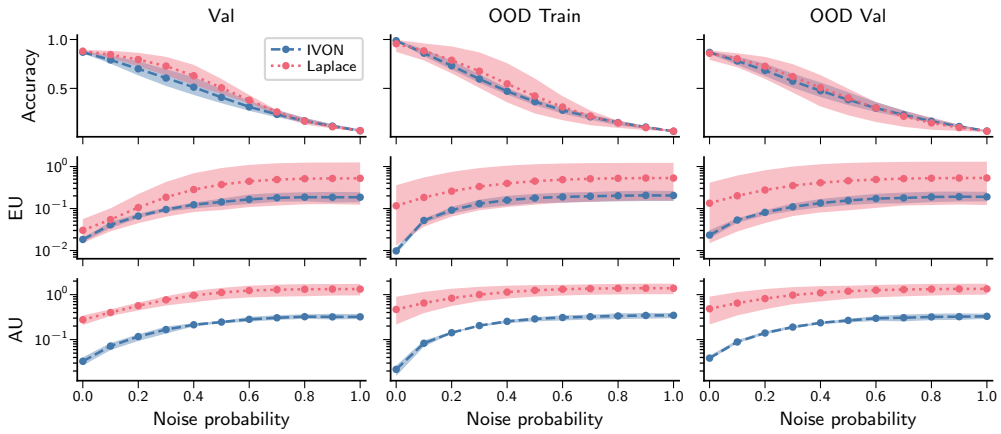


Figure 11: **Full split-wise evaluation under y_i-token corruption.** For fixed models trained at $(T = 256, \alpha = 0.8, K = 32)$, replacing each completed example's y_i token with a uniformly sampled token with probability p_{noise} reduces mean accuracy and increases mean EU and AU for IVON and Last-Layer Laplace. Curves are arithmetic seed means, and bands span empirical 2.5th–97.5th percentiles across seeds, not confidence intervals; IVON uses seeds 0, 1, and 3, Laplace uses seeds 0–3, and both use 64 posterior samples.

B Extended Related Works

Grokking and In-Context Learning (ICL) are training-inference phenomena that expose a common theme in deep learning: generalization can emerge abruptly, late, and in ways not well predicted by early optimization signals. Grokking, originally characterized on small algorithmic datasets (notably modular arithmetic), is a delayed generalization regime where test performance remains poor long after training performance saturates, and then transitions sharply to near-perfect generalization (Power et al., 2022). ICL, popularized by large transformer language models (e.g., GPT-3), is the ability to adapt to a task at inference time by conditioning on a prompt containing input-output examples, without gradient-based parameter updates (Brown et al., 2020). Explanations of ICL increasingly characterize transformers as implementing recognizable statistical estimators or learning procedures in their forward pass, including ridge regression, gradient descent, and in certain regimes

Bayesian estimators (Akyurek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023; Xie et al., 2025). A Bayesian framing provides a natural language for connecting these threads: pretraining can be interpreted as learning an implicit prior (or an implicit family of priors) over tasks/concepts, while ICL corresponds to approximate Bayesian inference (or Bayesian model averaging) over latent task variables given the prompt (Xie et al., 2022; Zhang et al., 2025). In parallel, BDL treats uncertainty as first-class, via priors and posteriors over parameters and predictive distributions, and offers tools that make uncertainty trajectories during training and inference measurable (Agostinelli et al., 2026; Graves, 2011; Blundell et al., 2015; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Maddox et al., 2019; Ritter et al., 2018).

In-context learning. While the theoretical understanding of ICL is still developing, a growing body of work has empirically studied various aspects of ICL, including calibration and uncertainty measures under prompting. Zhang et al. (2024) systematically study calibration under ICL across NLP tasks and report that increasing the number of ICL examples can initially worsen miscalibration before later improving calibration, and that miscalibration is prominent in low-shot settings. Ling et al. (2024) propose a formulation that decomposes predictive uncertainty in ICL into epistemic and aleatoric components and introduce estimation methods aimed at plug-and-play uncertainty decomposition for ICL. Wang et al. (2025) study long-context ICL and report that adding more in-context examples can reduce total uncertainty and epistemic uncertainty, while also analyzing confidence evolution across layers.

Grokking. Empirical studies of grokking have primarily characterized loss and accuracy dynamics, changes in representations, and mechanistic interpretations of the phenomenon. Power et al. (2022) established the canonical grokking setting on small modular arithmetic tasks, and observed that generalization emerges suddenly after a long period of overfitting. Liu et al. (2022) provided extensive phase-diagram evidence and identified multiple learning phases, emphasizing structured representation emergence as a driver of generalization. Mohamadi et al. (2024) focus specifically on modular addition and propose a theoretical explanation based on the Neural Tangent Kernel (NTK) regime, showing that the transition to generalization corresponding to departure from the kernel regime happens only after initial overfitting. A subset of grokking-adjacent work introduces information-theoretic progress measures, which is conceptually close to uncertainty quantification but not identical to predictive uncertainty. Clauw et al. (2024) analyze higher-order mutual information (synergy/redundancy) among neurons through training and report distinct phases that can anticipate grokking, framing grokking as an emergent phase transition caused by synergistic interactions. Golechha (2024) investigates grokking on real-world classification tasks and proposes progress measures such as activation sparsity, absolute weight entropy, and approximate local circuit complexity that correlate with grokking more strongly than weight norms in that study.

Difference from prior work. Closer to our work is the work of He et al. (2024), which extended the classic grokking setting to a larger class of modular linear functions, allowing for a more detailed analysis of the learning dynamics not only in-distribution but also out-of-distribution. While we build on the same experimental set up, our focus is on the Bayesian learning dynamics and uncertainty trajectories during grokking, rather than characterizing the phases of learning or the representational changes. We identify conditions under which the model generalizes in- and out-of-domain and analyze how its uncertainty evolves as it transitions from overfitting to generalization.

C Task Construction and Inference Methods

C.1 Task distribution and sequence construction

Each task is a coefficient pair $\tau = (a, b) \in \{0, \dots, p-1\}^2$, and each point is an input pair (x, y) from the same Cartesian product. Their deterministic target is

$$z = f_\tau(x, y) = (ax + by) \bmod p.$$

The implementation enumerates all p^2 input pairs, shuffles them once using the data seed, and assigns the first $\lfloor \alpha p^2 \rfloor$ pairs to the training-point split and the remainder to the validation-point split. It samples the base in-distribution task pairs uniformly without replacement. Out-of-distribution tasks are sampled from the complement of the final augmented in-distribution task set, so the ID/OOD distinction concerns the coefficient pair while the train/validation distinction concerns the input pair.

Parallelogram task augmentation. For every sampled base task (a, b), the generator draws two coordinates c, d that are distinct from a, b, and one another, and adds (a, d), (c, b), and (c, d). The new pairs are accepted only if none is already in the accumulated task set. Thus the four coefficient pairs form the corners of an axis-aligned parallelogram in task space and the procedure expands n_{task} base tasks to $T = 4n_{\text{task}}$ related tasks; this is the task-sharing construction of He et al. (2024).

Task-balanced batching. For a fixed point index, a dataset item contains the corresponding (x, y, z) row for every task. The collator stacks r such items into a tensor of shape $T \times r \times 3$ and, during ID training, applies one shared permutation π to the point axis for all tasks. It then flattens the task and point axes and partitions the result into sequences of K triplets. In the reported grids, $r = BK/T$ is a multiple of K, so each sequence contains points from exactly one task, the effective batch contains B sequences, and corresponding sequences for different tasks receive the same ordered inputs. Validation and OOD collation omit the random permutation, making their ordering deterministic.

Model likelihood. For a sequence of K triplets, let

$$\boldsymbol{c}_i = (x_1, y_1, z_1, \ldots, x_{i-1}, y_{i-1}, z_{i-1}, x_i, y_i)$$

be the completed preceding triples together with the current input pair. The causal Transformer $f(\boldsymbol{\theta}, \boldsymbol{c}_i)$ induces a categorical likelihood $p(z_i \mid \boldsymbol{c}_i)$ over the p possible answers. Conditioning on all input pairs, the answer-sequence likelihood factorizes autoregressively as

$$p(z_{1:K} \mid x_{1:K}, y_{1:K}) = \prod_{i=1}^{K} p(z_i \mid \boldsymbol{c}_i).$$

For a training minibatch of B sequences, indexed by b, the implementation estimates the per-answer log-likelihood by

$$\widehat{\ell}_{\text{all}}(\mathcal{B}) = \frac{1}{BK} \sum_{b=1}^{B} \sum_{i=1}^{K} \log p(z_{b,i} \mid \boldsymbol{c}_{b,i}),$$

The concrete task-generation and sequence settings used in the experiments are collected in Table 5; all remaining architecture, optimization, posterior, sweep, and hardware configurations are given in Appendix E.

C.2 Improved Variational Online Newton (IVON)

IVON (Shen et al., 2024) is a scalable, second-order optimizer that makes variational inference tractable for large-scale deep neural networks. Instead of minimizing the empirical risk, IVON directly optimizes the Evidence Lower Bound (ELBO):

$$\mathcal{L}(q) = \lambda \mathbb{E}_{q(\boldsymbol{w})}[\bar{\ell}(\boldsymbol{w})] + \mathbb{D}_{KL}(q(\boldsymbol{w}) || p(\boldsymbol{w})) \tag{17}$$

where, $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \text{diag}(\boldsymbol{\Sigma})^2)$ is the diagonal Gaussian posterior approximation, $p(\boldsymbol{w})$ is the prior, $\bar{\ell}(\boldsymbol{w})$ is the empirical loss, and λ scales the objective.

Previous methods like VON (Khan et al., 2018) relied on a computationally expensive per-example Gauss-Newton Hessian estimate. IVON circumvents this by using the *reparameterization trick* to estimate the Hessian $\hat{\boldsymbol{h}}$ using only the standard minibatch gradient:

$$\hat{\boldsymbol{h}} \leftarrow \hat{\nabla}\bar{\ell}(\boldsymbol{w}) \odot \frac{\boldsymbol{w} - \boldsymbol{\mu}}{\boldsymbol{\Sigma}^2} \tag{18}$$

where $\odot$ is the element-wise product and $\boldsymbol{w} \sim q(\boldsymbol{w})$. By requiring only a single vector multiplication, IVON's computational cost matches Adam's.

To ensure the Hessian remains positive-definite, IVON applies a Riemannian gradient descent update:

$$\boldsymbol{h} \leftarrow \beta_2 \boldsymbol{h} + (1 - \beta_2)\hat{\boldsymbol{h}} + \frac{1}{2}(1 - \beta_2)^2 \frac{(\boldsymbol{h} - \hat{\boldsymbol{h}})^2}{\boldsymbol{h} + \delta} \tag{19}$$

Here, β_2 is the Hessian momentum, and $\delta > 0$ acts as a damping parameter corresponding to the prior precision (weight decay).

The mean $\boldsymbol{\mu}$ is updated via a preconditioned Newton-like step, and the standard deviation is tied to the Hessian via $\boldsymbol{\Sigma} = 1/\sqrt{\lambda(\boldsymbol{h} + \delta)}$. At inference, IVON quantifies uncertainty through Bayesian Model Averaging (BMA) by sampling $\boldsymbol{w}^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\Sigma})^2)$.

C.3 Laplace approximation (LA)

Laplace approximation (Daxberger et al., 2021) is a *post-hoc* framework that converts a deterministically trained neural network into a Bayesian model for uncertainty quantification, without modifying the training objective.

LA interprets the final weights trained with L_2 regularization as a Maximum A Posteriori (MAP) estimate, $\boldsymbol{w}_{\text{MAP}}$, obtained by minimizing the regularized negative log-posterior:

$$\boldsymbol{w}_{\text{MAP}} = \arg\min_{\boldsymbol{w}} \left(-\sum_{n=1}^{N} \log p(y_n \mid f_{\boldsymbol{w}}(\boldsymbol{x}_n)) - \log p(\boldsymbol{w}) \right) \tag{20}$$

where $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \tau^2 \boldsymbol{I}_d)$ is the isotropic Gaussian prior induced by weight decay.

LA constructs a local Gaussian approximation of the intractable true posterior centered at the MAP solution:

$$p(\boldsymbol{w} \mid \mathcal{D}) \approx \mathcal{N}(\boldsymbol{w}; \boldsymbol{w}_{\text{MAP}}, \boldsymbol{\Sigma}) \tag{21}$$

The covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{H}^{-1}$ is the inverse of the Hessian $\boldsymbol{H}$, derived via a second-order Taylor expansion evaluated at $\boldsymbol{w}_{\text{MAP}}$.

Computing and inverting the full exact Hessian $\boldsymbol{H} \in \mathbb{R}^{d \times d}$ is computationally infeasible and often yields an indefinite matrix. Practical LA implementations resolve this through three core approximations:

1. Fisher Information Matrix (FIM): Replaces the exact Hessian to guarantee positive semi-definiteness.
2. Last-Layer Laplace (LLLA): Restricts Bayesian inference to the final linear layer while keeping preceding layers fixed at $\boldsymbol{w}_{\text{MAP}}$, drastically reducing dimensionality.
3. Kronecker-Factored Approximate Curvature (KFAC): Approximates the Fisher matrix as a Kronecker product $F \approx A \otimes B$ (input and pre-activation gradient covariances) for efficient inversion.

The predictive distribution for a new data point $\boldsymbol{x}^*$ is computed by marginalizing over this approximate posterior:

$$p(y \mid \boldsymbol{x}^*, \mathcal{D}) \approx \int p(y \mid f_{\boldsymbol{w}}(\boldsymbol{x}^*)) \mathcal{N}(\boldsymbol{w}; \boldsymbol{w}_{\text{MAP}}, \boldsymbol{\Sigma}) \, \mathrm{d}\boldsymbol{w} \tag{22}$$

In practice, this integral is approximated via Monte Carlo sampling, producing a distribution of outputs to quantify epistemic uncertainty.

D A Simplified Model for Grokking under Variational Inference

This section rewrites the derivation in a self-contained form for a linear-Gaussian teacher-student model. The goal is to isolate the mechanism that produces distinct training and generalization timescales under Gaussian variational inference, while keeping all quantities analytically tractable. Note that while the proof style and some of the intermediate results are similar to those in Levi et al. (2024), the final results and the interpretation are different, as we focus on the variational learning

dynamics rather than the gradient flow of the population risk. Furthermore, the analysis in Levi et al. (2024) is mostly focused on loss and accuracy dynamics, while we make a specific focus on the uncertainty dynamics and how they relate to the generalization transition, which is a novel perspective on the grokking phenomenon.

D.1 Bayesian linear regression

Definition D.1 (Linear-Gaussian model)**.** Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be the design matrix, let $\boldsymbol{w} \in \mathbb{R}^d$ be the parameter vector, and let $\boldsymbol{y} \in \mathbb{R}^n$ be the observed responses. We consider the Bayesian linear model

$$p(\boldsymbol{y} \mid \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}_n),$$
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \tau^2 \boldsymbol{I}_d),$$

with likelihood variance $\sigma^2 > 0$ and prior variance $\tau^2 > 0$. When needed, we also assume the true data-generating process is linear-Gaussian:

$$y = \boldsymbol{x}^\top \boldsymbol{w}_\star + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathbb{E}[\varepsilon^2] = \sigma_\varepsilon^2,$$

with population covariance $\boldsymbol{\Sigma}_x \equiv \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$.

Definition D.2 (Posterior precision and linear term)**.** For the model above, define

$$\boldsymbol{A} \equiv \frac{\boldsymbol{X}^\top \boldsymbol{X}}{\sigma^2} + \frac{\boldsymbol{I}_d}{\tau^2},$$
$$\boldsymbol{b} \equiv \frac{\boldsymbol{X}^\top \boldsymbol{y}}{\sigma^2}.$$

Since $\tau^2 > 0$, the matrix $\boldsymbol{A}$ is symmetric positive definite.

The posterior distribution for this model is the following Gaussian:

$$p(\boldsymbol{w} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1},$$
$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{b}.$$

D.2 Gaussian variational inference as a Wasserstein flow

Definition D.3 (Gaussian variational family and Bures-Wasserstein flow)**.** We approximate the posterior by a time-dependent Gaussian

$$q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t),$$

and study the Bures-Wasserstein gradient flow of the variational objective (Lambert et al., 2022). In the Gaussian family, the flow can be written as

$$\dot{\boldsymbol{\mu}}_t = -\mathbb{E}_{q_t}\left[\nabla_{\boldsymbol{w}_t} \log \frac{q_t(\boldsymbol{w}_t)}{p(\boldsymbol{y}, \boldsymbol{w}_t)}\right],$$
$$\dot{\boldsymbol{\Sigma}}_t = -\mathbb{E}_{q_t}\left[\left(\nabla_{\boldsymbol{w}_t} \log \frac{q_t(\boldsymbol{w}_t)}{p(\boldsymbol{y}, \boldsymbol{w}_t)}\right)(\boldsymbol{w}_t - \boldsymbol{\mu}_t)^\top + (\boldsymbol{w}_t - \boldsymbol{\mu}_t)\left(\nabla_{\boldsymbol{w}_t} \log \frac{q_t(\boldsymbol{w}_t)}{p(\boldsymbol{y}, \boldsymbol{w}_t)}\right)^\top\right].$$

Proposition D.1 (Closed ODEs for the variational mean and covariance)**.** *The Gaussian variational parameters obey*

$$\dot{\boldsymbol{\mu}}_t = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{\mu}_t, \tag{23}$$
$$\dot{\boldsymbol{\Sigma}}_t = 2\boldsymbol{I}_d - \boldsymbol{A}\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t\boldsymbol{A}. \tag{24}$$

Proof. We have

$$\log q_t(\boldsymbol{w}) = -\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_t)^\top \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_t) + \text{const},$$

so

$$\nabla_{\boldsymbol{w}} \log q_t(\boldsymbol{w}) = -\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_t),$$
$$\nabla_{\boldsymbol{w}} \log p(\boldsymbol{y}, \boldsymbol{w}) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{w}.$$

Subtracting these two gradients yields

$$-\nabla_{\boldsymbol{w}} \log \frac{q_t(\boldsymbol{w})}{p(\boldsymbol{y}, \boldsymbol{w})} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{w} + \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_t).$$

Taking the expectation of this identity under q_t and using $\mathbb{E}_{q_t}[\boldsymbol{w}] = \boldsymbol{\mu}_t$ gives

$$-\mathbb{E}_{q_t}\left[\nabla_{\boldsymbol{w}} \log \frac{q_t(\boldsymbol{w})}{p(\boldsymbol{y}, \boldsymbol{w})}\right] = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{\mu}_t,$$

which proves Equation (23).

For the covariance dynamics, set $\boldsymbol{d} = \boldsymbol{w} - \boldsymbol{\mu}_t$ so that $\mathbb{E}_{q_t}[\boldsymbol{d}] = \boldsymbol{0}$ and $\mathbb{E}_{q_t}[\boldsymbol{d}\boldsymbol{d}^\top] = \boldsymbol{\Sigma}_t$. Then

$$\begin{aligned}\mathbb{E}_{q_t}\left[\left(-\nabla_{\boldsymbol{w}} \log \frac{q_t(\boldsymbol{w})}{p(\boldsymbol{y}, \boldsymbol{w})}\right)\boldsymbol{d}^\top\right] &= \mathbb{E}_{q_t}\left[(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{w})\boldsymbol{d}^\top\right] + \mathbb{E}_{q_t}\left[\boldsymbol{\Sigma}_t^{-1}\boldsymbol{d}\boldsymbol{d}^\top\right] \\ &= -\boldsymbol{A}\boldsymbol{\Sigma}_t + \boldsymbol{I}_d.\end{aligned}$$

Indeed, the first term becomes

$$\mathbb{E}_{q_t}\left[(\boldsymbol{b} - \boldsymbol{A}(\boldsymbol{\mu}_t + \boldsymbol{d}))\boldsymbol{d}^\top\right] = -\boldsymbol{A}\mathbb{E}_{q_t}[\boldsymbol{d}\boldsymbol{d}^\top] = -\boldsymbol{A}\boldsymbol{\Sigma}_t,$$

and the second simplifies to

$$\mathbb{E}_{q_t}\left[\boldsymbol{\Sigma}_t^{-1}\boldsymbol{d}\boldsymbol{d}^\top\right] = \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_t = \boldsymbol{I}_d.$$

Plugging this identity and its transpose into the Bures–Wasserstein flow defined above yields

$$\dot{\boldsymbol{\Sigma}}_t = (\boldsymbol{I}_d - \boldsymbol{A}\boldsymbol{\Sigma}_t) + (\boldsymbol{I}_d - \boldsymbol{A}\boldsymbol{\Sigma}_t)^\top = 2\boldsymbol{I}_d - \boldsymbol{A}\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t\boldsymbol{A},$$

because $\boldsymbol{A}$ is symmetric. □

Theorem D.1 (Explicit variational trajectory and convergence)**.** *Let $\boldsymbol{\mu}_\infty \equiv \boldsymbol{A}^{-1}\boldsymbol{b}$ and $\boldsymbol{\Sigma}_\infty \equiv \boldsymbol{A}^{-1}$. Then the solution of Equations (23) and (24) is*

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_\infty + e^{-\boldsymbol{A}t}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_\infty), \tag{25}$$
$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_\infty + e^{-\boldsymbol{A}t}(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_\infty)e^{-\boldsymbol{A}t}. \tag{26}$$

In particular, $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ converges exponentially fast to the exact posterior parameters displayed above.

Proof. The fixed point of Equation (23) is $\boldsymbol{\mu}_\infty = \boldsymbol{A}^{-1}\boldsymbol{b}$. Writing $\delta\boldsymbol{\mu}_t \equiv \boldsymbol{\mu}_t - \boldsymbol{\mu}_\infty$, we obtain

$$\dot{\delta\boldsymbol{\mu}}_t = -\boldsymbol{A}\delta\boldsymbol{\mu}_t,$$

hence $\delta\boldsymbol{\mu}_t = e^{-\boldsymbol{A}t}\delta\boldsymbol{\mu}_0$, which proves Equation (25).

Likewise, $\boldsymbol{\Sigma}_\infty = \boldsymbol{A}^{-1}$ solves the stationary equation $2\boldsymbol{I}_d - \boldsymbol{A}\boldsymbol{\Sigma}_\infty - \boldsymbol{\Sigma}_\infty\boldsymbol{A} = \boldsymbol{0}$. Set $\Delta_t \equiv \boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_\infty$. Then

$$\dot{\Delta}_t = -\boldsymbol{A}\Delta_t - \Delta_t\boldsymbol{A}.$$

Since $\boldsymbol{A}$ is symmetric, the function $e^{-\boldsymbol{A}t}\Delta_0 e^{-\boldsymbol{A}t}$ satisfies the same ODE and initial condition, proving Equation (26). Because $\boldsymbol{A} \succ 0$, the matrix exponential decays exponentially, so $\boldsymbol{\mu}_t \to \boldsymbol{\mu}_\infty$ and $\boldsymbol{\Sigma}_t \to \boldsymbol{\Sigma}_\infty$. These limits are exactly the posterior mean and covariance. □

D.3 Empirical and generalization risks

Definition D.4 (Empirical and population risks)**.** For a deterministic predictor $\boldsymbol{w}$, define the squared-loss risks

$$\mathcal{R}_{\text{emp}}(\boldsymbol{w}) = \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2,$$
$$\mathcal{R}_{\text{gen}}(\boldsymbol{w}) = \frac{1}{2}\mathbb{E}_{p(\boldsymbol{x},y)}(y - \boldsymbol{x}^\top \boldsymbol{w})^2.$$

Along the Gaussian variational trajectory, we are interested in the expected risks

$$\mathcal{L}_{\text{emp}}(t) \equiv \mathbb{E}_{q_t}\mathcal{R}_{\text{emp}}(\boldsymbol{w}),$$
$$\mathcal{L}_{\text{gen}}(t) \equiv \mathbb{E}_{q_t}\mathcal{R}_{\text{gen}}(\boldsymbol{w}).$$

Proposition D.2 (Exact risks under a Gaussian predictor)**.** *For any Gaussian predictor $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$,*

$$\mathcal{L}_{emp}(t) = \frac{1}{2n}\left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_t\|^2 + \operatorname{Tr}\Big(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\Sigma}_t\Big)\right),$$
$$\mathcal{L}_{gen}(t) = \frac{1}{2}\left((\boldsymbol{\mu}_t - \boldsymbol{w}_\star)^\top \boldsymbol{\Sigma}_x(\boldsymbol{\mu}_t - \boldsymbol{w}_\star) + \operatorname{Tr}(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_t) + \sigma_\varepsilon^2\right).$$

Proof. Let $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. Then

$$\mathbb{E}_{q_t}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_t\|^2 + \operatorname{Tr}\Big(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\Sigma}_t\Big),$$

which proves the empirical-risk identity. For the population risk, write $y = \boldsymbol{x}^\top \boldsymbol{w}_\star + \varepsilon$ and use independence of ε, $\boldsymbol{x}$, and $\boldsymbol{w}$:

$$\begin{aligned}\mathbb{E}(y - \boldsymbol{x}^\top \boldsymbol{w})^2 &= \mathbb{E}\big[\boldsymbol{x}^\top(\boldsymbol{w}_\star - \boldsymbol{w})\big]^2 + \sigma_\varepsilon^2 \\ &= \mathbb{E}_{\boldsymbol{x}}\left[(\boldsymbol{w}_\star - \boldsymbol{\mu}_t)^\top \boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{w}_\star - \boldsymbol{\mu}_t)\right] + \mathbb{E}_{\boldsymbol{x}}\operatorname{Tr}\big(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}_t\big) + \sigma_\varepsilon^2 \\ &= (\boldsymbol{\mu}_t - \boldsymbol{w}_\star)^\top \boldsymbol{\Sigma}_x(\boldsymbol{\mu}_t - \boldsymbol{w}_\star) + \operatorname{Tr}(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_t) + \sigma_\varepsilon^2.\end{aligned}$$

Multiplying by $1/2$ gives the population-risk identity. □

Corollary D.1 (Exact risk decomposition along the VI flow)**.** *Define*

$$\boldsymbol{E}_t \equiv e^{-\boldsymbol{A}t}, \qquad \delta\boldsymbol{\mu}_0 \equiv \boldsymbol{\mu}_0 - \boldsymbol{\mu}_\infty, \qquad \Delta\boldsymbol{\Sigma}_0 \equiv \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_\infty.$$

Let $\boldsymbol{r}_\infty \equiv \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_\infty$ and $\boldsymbol{b}_\infty \equiv \boldsymbol{\mu}_\infty - \boldsymbol{w}_\star$. Then

$$\mathcal{L}_{emp}(t) = \mathcal{L}_{emp,\infty} - \frac{1}{n}\boldsymbol{r}_\infty^\top \boldsymbol{X}\boldsymbol{E}_t\delta\boldsymbol{\mu}_0 + \frac{1}{2n}\delta\boldsymbol{\mu}_0^\top \boldsymbol{E}_t\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{E}_t\delta\boldsymbol{\mu}_0 + \frac{1}{2n}\operatorname{Tr}\Big(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{E}_t\Delta\boldsymbol{\Sigma}_0\boldsymbol{E}_t\Big),$$
$$\mathcal{L}_{gen}(t) = \mathcal{L}_{gen,\infty} + \boldsymbol{b}_\infty^\top \boldsymbol{\Sigma}_x\boldsymbol{E}_t\delta\boldsymbol{\mu}_0 + \frac{1}{2}\delta\boldsymbol{\mu}_0^\top \boldsymbol{E}_t\boldsymbol{\Sigma}_x\boldsymbol{E}_t\delta\boldsymbol{\mu}_0 + \frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}_x\boldsymbol{E}_t\Delta\boldsymbol{\Sigma}_0\boldsymbol{E}_t),$$

where

$$\mathcal{L}_{emp,\infty} \equiv \frac{1}{2n}\left(\|\boldsymbol{r}_\infty\|^2 + \operatorname{Tr}\Big(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\Sigma}_\infty\Big)\right),$$
$$\mathcal{L}_{gen,\infty} \equiv \frac{1}{2}\left(\boldsymbol{b}_\infty^\top \boldsymbol{\Sigma}_x\boldsymbol{b}_\infty + \operatorname{Tr}(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_\infty) + \sigma_\varepsilon^2\right).$$

Proof. Substitute Equations (25) and (26) into the risk formulas of Proposition D.2 and expand the resulting quadratic forms. The empirical and generalization fixed-point terms are the parts independent of $\boldsymbol{E}_t$; the linear and quadratic time-dependent contributions are exactly those displayed above. □

Corollary D.2 (Matched interpolation regime)**.** *Assume $\boldsymbol{\Sigma}_x = \boldsymbol{I}_d$ and that the stationary predictor is well-specified in the sense that*

$$\boldsymbol{\mu}_\infty = \boldsymbol{w}_\star, \qquad \boldsymbol{r}_\infty = \boldsymbol{0}.$$

Define the excess risks

$$\widetilde{\mathcal{L}}_{emp}(t) \equiv \mathcal{L}_{emp}(t) - \mathcal{L}_{emp,\infty},$$
$$\widetilde{\mathcal{L}}_{gen}(t) \equiv \mathcal{L}_{gen}(t) - \mathcal{L}_{gen,\infty}.$$

Then

$$\widetilde{\mathcal{L}}_{emp}(t) = \frac{1}{2n}\delta\boldsymbol{\mu}_0^\top \boldsymbol{E}_t \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{E}_t \delta\boldsymbol{\mu}_0 + \frac{1}{2n}\operatorname{Tr}\Big(\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{E}_t \Delta\boldsymbol{\Sigma}_0 \boldsymbol{E}_t\Big),$$
$$\widetilde{\mathcal{L}}_{gen}(t) = \frac{1}{2}\delta\boldsymbol{\mu}_0^\top \boldsymbol{E}_t^2 \delta\boldsymbol{\mu}_0 + \frac{1}{2}\operatorname{Tr}(\boldsymbol{E}_t \Delta\boldsymbol{\Sigma}_0 \boldsymbol{E}_t).$$

Proof. In the preceding exact risk decomposition, the linear terms vanish because $\boldsymbol{r}_\infty = \boldsymbol{0}$ and $\boldsymbol{b}_\infty = \boldsymbol{0}$, while $\boldsymbol{\Sigma}_x = \boldsymbol{I}_d$ simplifies the quadratic and trace terms. □

D.4 Spectral reduction and random-matrix asymptotics

Lemma D.1 (Spectral form of the excess risks)**.** *Let*

$$\widehat{\boldsymbol{\Sigma}} \equiv \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U}\operatorname{diag}(\nu_1, \ldots, \nu_d)\boldsymbol{U}^\top.$$

Then

$$\boldsymbol{A} = \boldsymbol{U}\operatorname{diag}(a_1, \ldots, a_d)\boldsymbol{U}^\top, \qquad a_i = \frac{n}{\sigma^2}\nu_i + \frac{1}{\tau^2}.$$

Writing

$$\bar{\boldsymbol{d}}_0 \equiv \boldsymbol{U}^\top \delta\boldsymbol{\mu}_0, \qquad \bar{\boldsymbol{S}}_0 \equiv \boldsymbol{U}^\top \Delta\boldsymbol{\Sigma}_0 \boldsymbol{U},$$

the excess risks from Corollary D.2 become

$$\widetilde{\mathcal{L}}_{emp}(t) = \frac{1}{2}\sum_{i=1}^{d} \nu_i \big(\bar{d}_{0,i}^2 + (\bar{\boldsymbol{S}}_0)_{ii}\big) e^{-2a_i t}, \tag{27}$$

$$\widetilde{\mathcal{L}}_{gen}(t) = \frac{1}{2}\sum_{i=1}^{d} \big(\bar{d}_{0,i}^2 + (\bar{\boldsymbol{S}}_0)_{ii}\big) e^{-2a_i t}. \tag{28}$$

Proof. $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{A}$ are diagonalized by the same orthogonal matrix $\boldsymbol{U}$. In that basis,

$$\boldsymbol{E}_t = e^{-\boldsymbol{A}t} = \boldsymbol{U}\operatorname{diag}(e^{-a_1 t}, \ldots, e^{-a_d t})\boldsymbol{U}^\top.$$

Substituting this into the matched-interpolation formulas immediately yields the stated sums. □

Assumption D.1 (Self-averaging initialization)**.** In the large-system limit, assume that the initial mode weights are asymptotically uncorrelated with the eigenbasis of $\widehat{\boldsymbol{\Sigma}}$, so that

$$\bar{d}_{0,i}^2 + (\bar{\boldsymbol{S}}_0)_{ii} \approx \frac{\|\delta\boldsymbol{\mu}_0\|^2 + \operatorname{Tr}(\Delta\boldsymbol{\Sigma}_0)}{d}$$

uniformly over i.

Proposition D.3 (Self-averaged loss formulas)**.** *Under Assumption D.1, if*

$$C_0 \equiv \frac{1}{2}\left(\|\delta\boldsymbol{\mu}_0\|^2 + \operatorname{Tr}(\Delta\boldsymbol{\Sigma}_0)\right),$$

then

$$\widetilde{\mathcal{L}}_{emp}(t) \approx C_0 \frac{1}{d}\sum_{i=1}^{d} \nu_i \exp\left[-2\left(\frac{n}{\sigma^2}\nu_i + \frac{1}{\tau^2}\right)t\right],$$

$$\widetilde{\mathcal{L}}_{gen}(t) \approx C_0 \frac{1}{d}\sum_{i=1}^{d} \exp\left[-2\left(\frac{n}{\sigma^2}\nu_i + \frac{1}{\tau^2}\right)t\right].$$

If, moreover, $d, n \to \infty$ with aspect ratio $\lambda = d/n < 1$, then

$$\widetilde{\mathcal{L}}_{emp}(t) \approx C_0 \, \mathbb{E}_{\nu \sim \mathrm{MP}(\lambda)} \left[\nu \exp\left(-2 \left(\frac{n}{\sigma^2} \nu + \frac{1}{\tau^2} \right) t \right) \right], \tag{29}$$

$$\widetilde{\mathcal{L}}_{gen}(t) \approx C_0 \, \mathbb{E}_{\nu \sim \mathrm{MP}(\lambda)} \left[\exp\left(-2 \left(\frac{n}{\sigma^2} \nu + \frac{1}{\tau^2} \right) t \right) \right]. \tag{30}$$

Proof. Replace the mode-dependent prefactors in Equations (27) and (28) by the common self-averaged value from Assumption D.1. This yields the finite sums. The Marchenko-Pastur formulas follow by replacing the empirical spectral average with its large-d, n limit. □

Theorem D.2 (Late-time asymptotics and grokking delay)**.** *Assume the setting of Proposition D.3 with $\lambda < 1$, and let*

$$\nu_\pm = (1 \pm \sqrt{\lambda})^2, \qquad \kappa = 2 \left(\frac{n}{\sigma^2} (1 - \sqrt{\lambda})^2 + \frac{1}{\tau^2} \right).$$

Then, as $t \to \infty$,

$$\widetilde{\mathcal{L}}_{emp}(t) \sim K_{emp} \, t^{-3/2} e^{-\kappa t}, \tag{31}$$

$$\widetilde{\mathcal{L}}_{gen}(t) \sim K_{gen} \, t^{-3/2} e^{-\kappa t}, \tag{32}$$

with prefactor ratio

$$\frac{K_{gen}}{K_{emp}} = \frac{1}{\nu_-} = \frac{1}{(1 - \sqrt{\lambda})^2}. \tag{33}$$

If the grokking time is defined by the 95% accuracy threshold

$$\widetilde{\mathcal{L}}(t^\star) = \frac{\epsilon}{8},$$

then the leading train-generalization delay is

$$\Delta t_{grok} \equiv t^\star_{gen} - t^\star_{emp} \simeq \frac{\log\left(\frac{1}{1 - \sqrt{\lambda}} \right)}{\frac{n}{\sigma^2} (1 - \sqrt{\lambda})^2 + \frac{1}{\tau^2}}. \tag{34}$$

Proof. The Marchenko-Pastur density on $[\nu_-, \nu_+]$ behaves near its lower edge as

$$p_{\mathrm{MP}}(\nu) \propto \sqrt{\nu - \nu_-}.$$

Write $\nu = \nu_- + u$. In Equations (29) and (30), the exponential factor is $e^{-\kappa t} e^{-2(n/\sigma^2) u t}$. Hence both integrals are governed by the same edge contribution

$$e^{-\kappa t} \int_0^\infty u^{1/2} e^{-2(n/\sigma^2) u t} \mathrm{d}u \propto t^{-3/2} e^{-\kappa t},$$

which proves Equations (31) and (32). The empirical loss contains an extra factor $\nu = \nu_- + u$, which contributes only the multiplicative constant ν_- at leading order; this yields Equation (33).

To obtain the delay, solve $K t^{-3/2} e^{-\kappa t} = \epsilon/8$ asymptotically for t. The subleading $t^{-3/2}$ correction is the same for the empirical and generalization losses, so the leading difference between the two solutions is

$$\Delta t_{\mathrm{grok}} \simeq \frac{\log(K_{\mathrm{gen}}/K_{\mathrm{emp}})}{\kappa}.$$

Using Equation (33) and the definition of κ gives Equation (34). □

D.5 Accuracy and uncertainty of the sampled predictor

Proposition D.4 (Exact accuracy of the sampled VI predictor)**.** *Let $\boldsymbol{w} \sim q_t$. For a fixed datapoint $(\boldsymbol{x}, y)$, define*

$$m_t(\boldsymbol{x}, y) \equiv y - \boldsymbol{x}^\top \boldsymbol{\mu}_t, \qquad v_t(\boldsymbol{x}) \equiv \boldsymbol{x}^\top \boldsymbol{\Sigma}_t \boldsymbol{x}.$$

Then the residual

$$r_t(\boldsymbol{x}, y; \boldsymbol{w}) \equiv y - \boldsymbol{x}^\top \boldsymbol{w}$$

is conditionally Gaussian under q_t, and for any tolerance $\epsilon > 0$ the empirical and population accuracies are

$$\mathcal{A}_{emp}(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\Phi\left(\frac{\sqrt{\epsilon} - m_i(t)}{\sqrt{v_i(t)}} \right) - \Phi\left(\frac{-\sqrt{\epsilon} - m_i(t)}{\sqrt{v_i(t)}} \right) \right],$$

$$\mathcal{A}_{gen}(t) = \mathbb{E}_{p(\boldsymbol{x})} \left[\Phi\left(\frac{\sqrt{\epsilon} - \boldsymbol{x}^\top(\boldsymbol{w}_\star - \boldsymbol{\mu}_t)}{\sqrt{\sigma_\varepsilon^2 + \boldsymbol{x}^\top \boldsymbol{\Sigma}_t \boldsymbol{x}}} \right) - \Phi\left(\frac{-\sqrt{\epsilon} - \boldsymbol{x}^\top(\boldsymbol{w}_\star - \boldsymbol{\mu}_t)}{\sqrt{\sigma_\varepsilon^2 + \boldsymbol{x}^\top \boldsymbol{\Sigma}_t \boldsymbol{x}}} \right) \right],$$

where $m_i(t) = y_i - \boldsymbol{x}_i^\top \boldsymbol{\mu}_t$ and $v_i(t) = \boldsymbol{x}_i^\top \boldsymbol{\Sigma}_t \boldsymbol{x}_i$, and Φ is the standard Gaussian CDF.

Proof. Since $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, the scalar random variable $\boldsymbol{x}^\top \boldsymbol{w}$ is Gaussian with mean $\boldsymbol{x}^\top \boldsymbol{\mu}_t$ and variance $\boldsymbol{x}^\top \boldsymbol{\Sigma}_t \boldsymbol{x}$. Therefore

$$r_t(\boldsymbol{x}, y; \boldsymbol{w}) \mid (\boldsymbol{x}, y) \sim \mathcal{N}(m_t(\boldsymbol{x}, y), v_t(\boldsymbol{x})).$$

The formulas follow by evaluating the Gaussian probability of the interval $[-\sqrt{\epsilon}, \sqrt{\epsilon}]$. For the generalization expression, use $y = \boldsymbol{x}^\top \boldsymbol{w}_\star + \varepsilon$ and absorb the label noise into the conditional variance. □

Proposition D.5 (Entropy decomposition of predictive uncertainty)**.** *Assume the predictive model*

$$y \mid \boldsymbol{x}, \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{w}, \sigma_\varepsilon^2).$$

Then the predictive entropy decomposes as

$$\mathbb{H}[y \mid \boldsymbol{x}, \mathcal{D}, t] = \mathbb{E}_{q_t(\boldsymbol{w})} \mathbb{H}[y \mid \boldsymbol{x}, \boldsymbol{w}] + \mathbb{I}_t(y; \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D}).$$

Moreover, the predictive distribution is

$$p_t(y \mid \boldsymbol{x}, \mathcal{D}) = \mathcal{N}\left(y; \boldsymbol{x}^\top \boldsymbol{\mu}_t, \sigma_\varepsilon^2 + \boldsymbol{x}^\top \boldsymbol{\Sigma}_t \boldsymbol{x}\right),$$

and the pointwise epistemic uncertainty is

$$EU(\boldsymbol{x}, t) = \frac{1}{2} \log\left(1 + \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma}_t \boldsymbol{x}}{\sigma_\varepsilon^2}\right). \tag{35}$$

Using the explicit covariance trajectory from Equation (26), this can also be written as

$$EU(\boldsymbol{x}, t) = \frac{1}{2} \log\left(1 + \frac{\boldsymbol{x}^\top \boldsymbol{\Sigma}_\infty \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{E}_t \Delta \boldsymbol{\Sigma}_0 \boldsymbol{E}_t \boldsymbol{x}}{\sigma_\varepsilon^2}\right).$$

Proof. The entropy decomposition is the standard identity $H(Y \mid X) = \mathbb{E}H(Y \mid X, W) + I(Y; W \mid X)$. Since both $p(y \mid \boldsymbol{x}, \boldsymbol{w})$ and $q_t(\boldsymbol{w})$ are Gaussian, marginalizing over $\boldsymbol{w}$ gives the predictive distribution above. The differential entropy of a Gaussian $\mathcal{N}(0, s^2)$ is $\frac{1}{2}\log(2\pi e s^2)$, so the aleatoric term is $\frac{1}{2}\log(2\pi e \sigma_\varepsilon^2)$ and the epistemic contribution is the difference between the predictive and aleatoric entropies, which yields Equation (35). □

Remark D.1 (Late-time accuracy closure). In the matched interpolation regime, the late-time residual is approximately centered and its variance is controlled by the excess risk:

$$\mathbb{E}[r_t^2] \approx 2\widetilde{\mathcal{L}}(t).$$

Approximating the residual by a single effective Gaussian therefore gives

$$\mathcal{A}_{\text{emp/gen}}(t) \approx \operatorname{erf}\left(\sqrt{\frac{\epsilon}{4\widetilde{\mathcal{L}}_{\text{emp/gen}}(t)}} \right).$$

The 95% grokking threshold used in Theorem D.2 then corresponds to $\widetilde{\mathcal{L}}(t^\star) = \epsilon/8$ because $\operatorname{erf}(\sqrt{2}) \approx 0.95$.

Lemma D.2 (Trace dynamics of the covariance)**.** *The covariance trace obeys*

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Tr}(\boldsymbol{\Sigma}_t) = 2\operatorname{Tr}(\boldsymbol{I}_d - \boldsymbol{A}\boldsymbol{\Sigma}_t).$$

In the eigenbasis of $\boldsymbol{A}$, if $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_\infty + \boldsymbol{E}_t\Delta\boldsymbol{\Sigma}_0\boldsymbol{E}_t$ and Δs_i denotes the i-th diagonal entry of $\boldsymbol{U}^\top\Delta\boldsymbol{\Sigma}_0\boldsymbol{U}$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Tr}(\boldsymbol{\Sigma}_t) = -2\sum_{i=1}^{d} a_i\Delta s_i e^{-2a_i t}.$$

Proof. Taking the trace of Equation (24) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Tr}(\boldsymbol{\Sigma}_t) = 2\operatorname{Tr}(\boldsymbol{I}_d) - \operatorname{Tr}(\boldsymbol{A}\boldsymbol{\Sigma}_t) - \operatorname{Tr}(\boldsymbol{\Sigma}_t\boldsymbol{A}) = 2\operatorname{Tr}(\boldsymbol{I}_d - \boldsymbol{A}\boldsymbol{\Sigma}_t),$$

using cyclicity of the trace. The spectral formula follows by substituting the explicit solution from Equation (26). □

Lemma D.3 (Population-averaged epistemic uncertainty)**.** *The exact population-averaged epistemic uncertainty is*

$$EU(t) \equiv \mathbb{E}_{p(\boldsymbol{x})}EU(\boldsymbol{x},t) = \frac{1}{2}\mathbb{E}_{p(\boldsymbol{x})}\log\left(1 + \frac{\boldsymbol{x}^\top\boldsymbol{\Sigma}_t\boldsymbol{x}}{\sigma_\varepsilon^2}\right).$$

Under isotropic Gaussian design $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and the same self-averaging closure used above,

$$EU(t) \approx \frac{1}{2}\log\left(1 + \frac{\operatorname{Tr}(\boldsymbol{\Sigma}_t)}{\sigma_\varepsilon^2}\right). \tag{36}$$

If $\boldsymbol{\Sigma}_0 \prec \boldsymbol{\Sigma}_\infty$, this approximate epistemic curve is monotonically increasing; if $\boldsymbol{\Sigma}_0 \succ \boldsymbol{\Sigma}_\infty$, it is monotonically decreasing; and if $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_\infty$, it is constant.

Proof. The exact identity is just the average of Equation (35). Under isotropic Gaussian design, the quadratic form $\boldsymbol{x}^\top\boldsymbol{\Sigma}_t\boldsymbol{x}$ concentrates around its trace in high dimension, yielding Equation (36). Because $u \mapsto \frac{1}{2}\log(1 + u/\sigma_\varepsilon^2)$ is increasing, the monotonicity of the approximate entropy curve matches the monotonicity of $\operatorname{Tr}(\boldsymbol{\Sigma}_t)$. The explicit solution Equation (26) shows that each spectral mode evolves as $a_i^{-1} + \Delta s_i e^{-2a_i t}$, so the sign of Δs_i determines whether that mode increases or decreases. If $\boldsymbol{\Sigma}_0 \prec \boldsymbol{\Sigma}_\infty$, then all $\Delta s_i < 0$ and Lemma D.2 implies that $\operatorname{Tr}(\boldsymbol{\Sigma}_t)$ increases monotonically; the opposite case is analogous, and equality gives a constant curve. □

Corollary D.3 (Isotropic initialization)**.** *If $\boldsymbol{\Sigma}_0 = s_0\boldsymbol{I}_d$, then*

$$\frac{1}{d}\operatorname{Tr}(\boldsymbol{\Sigma}_t) \approx \mathbb{E}_{\nu\sim\mathrm{MP}(\lambda)}\left[\frac{1}{a(\nu)} + \left(s_0 - \frac{1}{a(\nu)}\right)e^{-2a(\nu)t}\right], \qquad a(\nu) = \frac{n}{\sigma^2}\nu + \frac{1}{\tau^2}.$$

In particular, when $\boldsymbol{\Sigma}_0 = \boldsymbol{0}$,

$$\frac{1}{d}\operatorname{Tr}(\boldsymbol{\Sigma}_t) \approx \mathbb{E}_{\nu\sim\mathrm{MP}(\lambda)}\left[\frac{1 - e^{-2a(\nu)t}}{a(\nu)}\right].$$

Proof. Under isotropic initialization, $\Delta\boldsymbol{\Sigma}_0$ is diagonal in every eigenbasis, so the spectral representation of Equation (26) reduces to a scalar mode-by-mode formula. Averaging over the empirical spectrum and then passing to the Marchenko-Pastur limit gives the result. □

Proposition D.6 (Epistemic uncertainty of the threshold model)**.** *Define the thresholded success indicator*

$$Z_\epsilon(\boldsymbol{x}, y; \boldsymbol{w}) \equiv \mathbb{1}\{(y - \boldsymbol{x}^\top\boldsymbol{w})^2 \le \epsilon\}.$$

For fixed $(\boldsymbol{x}, y)$, the success probability under q_t is

$$p_t(\boldsymbol{x}, y) = \Phi\left(\frac{\sqrt{\epsilon} - m_t(\boldsymbol{x}, y)}{\sqrt{v_t(\boldsymbol{x})}}\right) - \Phi\left(\frac{-\sqrt{\epsilon} - m_t(\boldsymbol{x}, y)}{\sqrt{v_t(\boldsymbol{x})}}\right),$$

and

$$Z_\epsilon(\boldsymbol{x}, y; \boldsymbol{w}) \mid (\boldsymbol{x}, y, \mathcal{D}, t) \sim \text{Bernoulli}(p_t(\boldsymbol{x}, y)).$$

Hence the threshold-model epistemic uncertainty is exactly

$$EU_t(Z_\epsilon; \boldsymbol{w} \mid \boldsymbol{x}, y, \mathcal{D}) = h_2\big(p_t(\boldsymbol{x}, y)\big),$$

where $h_2(p) = -p \log p - (1-p)\log(1-p)$ *is the binary entropy. In the late-time self-averaging regime,*

$$p_t \approx \text{erf}\left(\sqrt{\frac{\epsilon}{4\widetilde{\mathcal{L}}(t)}}\right),$$

and therefore

$$EU_{\text{th}}(t) \approx h_2\left(\text{erf}\left(\sqrt{\frac{\epsilon}{4\widetilde{\mathcal{L}}(t)}}\right)\right). \tag{37}$$

Proof. The exact success probability is the Gaussian interval probability from Proposition D.4. Conditional on $\boldsymbol{w}$, the variable Z_ϵ is deterministic, so its conditional entropy vanishes. The predictive distribution is therefore Bernoulli with entropy $h_2(p_t(\boldsymbol{x}, y))$, and the full uncertainty is epistemic. The closure formula follows from the same effective-Gaussian approximation that relates the residual variance to the excess risk at late times. □

Remark D.2 (Threshold uncertainty versus grokking). The binary entropy $h_2(p)$ is maximized at $p = 1/2$, so the threshold-model epistemic uncertainty typically follows a non-monotone trajectory: it rises while predictions are ambiguous and then falls as the model becomes confidently correct. Using the late-time closure in the preceding proposition, the peak occurs when

$$\widetilde{\mathcal{L}}(t_{\text{peak}}) = \frac{\epsilon}{4(\text{erf}^{-1}(1/2))^2}, \tag{38}$$

and numerically

$$\frac{1}{4(\text{erf}^{-1}(1/2))^2} \approx 1.10.$$

Thus the peak occurs when the loss is still of order ϵ, which is much earlier than the 95% grokking threshold $\widetilde{\mathcal{L}}(t^\star) = \epsilon/8$. Thus the entropy peak and the grokking delay are distinct observables, even though both are controlled by the same slow edge-of-spectrum modes in the Marchenko-Pastur regime.

Theorem D.3 (Delay between threshold-uncertainty peaks)**.** *Assume the setting of Theorem D.2 and Proposition D.6, and suppose the empirical and generalization threshold-model uncertainty peaks occur in the late-time regime where Equations (31), (32) and (37) apply. Then*

$$\begin{aligned}\Delta t_{EU\,peak}(\lambda) &\equiv t_{\text{peak,gen}} - t_{\text{peak,emp}} \\ &\simeq \frac{\log\left(\frac{1}{1-\sqrt{\lambda}}\right)}{\frac{n}{\sigma^2}(1-\sqrt{\lambda})^2 + \frac{1}{\tau^2}},\end{aligned} \tag{39}$$

up to threshold-dependent subleading corrections.

Proof. For $\alpha \in \{\text{emp}, \text{gen}\}$, write

$$h_\alpha(t) \approx h_2\big(p_\alpha(t)\big), \qquad p_\alpha(t) \approx \text{erf}\left(\sqrt{\frac{\epsilon}{4\widetilde{\mathcal{L}}_\alpha(t)}}\right),$$

as in Equation (37). Since $h_2'(p) = \log\big(\frac{1-p}{p}\big)$, the binary entropy is uniquely maximized at $p = 1/2$. Moreover, the map

$$\widetilde{\mathcal{L}} \longmapsto \text{erf}\left(\sqrt{\frac{\epsilon}{4\widetilde{\mathcal{L}}}}\right)$$

is strictly decreasing in $\widetilde{\mathcal{L}}$, while in the late-time regime Equations (31) and (32) imply that each loss is strictly decreasing for large t. Hence each late-time uncertainty curve reaches its maximum exactly when $p_\alpha(t) = 1/2$, namely when

$$\widetilde{\mathcal{L}}_\alpha(t_{\text{peak},\alpha}) = c_{\text{peak}} \equiv \frac{\epsilon}{4(\text{erf}^{-1}(1/2))^2},$$

which is Equation (38).

Now use the late-time asymptotics from Theorem D.2. For each $\alpha \in \{\text{emp}, \text{gen}\}$,

$$\widetilde{\mathcal{L}}_\alpha(t) \sim K_\alpha\, t^{-3/2} e^{-\kappa t}.$$

Therefore the peak time satisfies

$$c_{\text{peak}} = K_\alpha\, t_{\text{peak},\alpha}^{-3/2} e^{-\kappa t_{\text{peak},\alpha}} \big(1 + o(1)\big).$$

Taking logarithms gives

$$\kappa t_{\text{peak},\alpha} = \log\left(\frac{K_\alpha}{c_{\text{peak}}}\right) - \frac{3}{2}\log t_{\text{peak},\alpha} + o(1).$$

Subtracting the empirical and generalization versions yields

$$\kappa \Delta t_{\text{EU peak}} \simeq \log\left(\frac{K_{\text{gen}}}{K_{\text{emp}}}\right) - \frac{3}{2}\log\left(\frac{t_{\text{peak,gen}}}{t_{\text{peak,emp}}}\right),$$

up to $o(1)$ terms. The logarithmic ratio of peak times is a subleading correction coming from the common $t^{-3/2}$ factor, exactly as in the proof of Theorem D.2. Thus the leading contribution is

$$\Delta t_{\text{EU peak}} \simeq \frac{\log(K_{\text{gen}}/K_{\text{emp}})}{\kappa},$$

with threshold dependence entering only through the discarded subleading terms. Finally, Equation (33) gives

$$\log\left(\frac{K_{\text{gen}}}{K_{\text{emp}}}\right) = -2\log\left(1 - \sqrt{\lambda}\right),$$

and

$$\kappa = 2\left(\frac{n}{\sigma^2}(1 - \sqrt{\lambda})^2 + \frac{1}{\tau^2}\right).$$

Substituting these expressions gives Equation (39). Thus the threshold-model uncertainty peak and the grokking delay are different observables, but near $\lambda \simeq 1$ they are governed by the same edge-of-spectrum slowdown. □

E Reproducibility

This appendix records the numerical configurations and experiment grids used for the reported results. The task-generation algorithm, batching construction, causal likelihood, and inference methods are described separately in Appendix C.

E.1 Task construction and sequence settings

Table 5 records the settings that determine the discrete task distribution and the causal sequences presented to every method. The modular vocabulary and the $K = 32$ construction are fixed across the reported experiments; only the base task count n_{task} and training-input fraction α vary over the experiment grids. Parallelogram augmentation converts each base task count to the effective count $T = 4n_{\text{task}}$, and the nominal batch size is fixed at 1024.

Table 5: Task-generation and sequence-construction choices shared across the main experiments.

Quantity	Value
Arithmetic rule	$z = (ax + by) \bmod 29$
Vocabulary size	29
Task dimension	2
Input-pair dimension	2
Input space size	$29^2 = 841$
Base in-distribution task count	n_{task}
Parallelogram augmentation	Enabled in all reported runs
Effective in-distribution task count	$4n_{\text{task}}$
Out-of-distribution task count	256
Training-input fraction α	$\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, depending on sweep
Constructed context length	$K = 32$ triplets: 31 in-context examples + query
Constructed token window	96 tokens
Causal model-input length	95 tokens
Nominal batch size	1024

Table 6: Shared architecture for MAP, IVON, and Laplace evaluations.

Quantity	Value
Architecture	Decoder-only transformer
Layers	6
Attention heads	4
Embedding width	512
Feed-forward width	2048
Vocabulary size	29
Maximum token window	96
Normalization	LayerNorm, $\epsilon = 10^{-5}$
Positional encoding	Rotary
Embedding tying	Enabled
Bias terms	Disabled
Numerical precision	Mixed bfloat16

E.2 Shared model architecture

All methods use the same decoder-only transformer. The backbone has 6 layers, 4 attention heads, embedding width 512, and feed-forward width 2048. The vocabulary size is 29 and the maximum token window is 96. The model uses LayerNorm with $\epsilon = 10^{-5}$, full rotary positional encoding, tied input and output embeddings, and no bias terms in attention, MLP, or output projections. All runs use mixed bfloat16 precision and one accelerator per run.

E.3 Training and evaluation protocol

The default optimization horizon is 100,000 update steps. For the largest base task setting, $n_{\text{task}} = 128$, the horizon is doubled to 200,000 steps. For every validation split we report accuracy, expected calibration error, mean log-likelihood, aleatoric uncertainty, and epistemic uncertainty. Accuracy, calibration, and uncertainty are reported both over all supervised answer positions in the sequence and on the final answer token only.

E.4 Method-specific settings

The MAP baseline is trained with AdamW using peak learning rate 1.5×10^{-4}, linear warmup for 1000 steps from 1% of the peak rate, and cosine decay to 10% of the peak rate. The momentum parameters are $(0.9, 0.98)$, the numerical stability constant is 10^{-8}, and the gradient norm is clipped at 1.0.

IVON uses learning rate 0.5, 2000 warmup steps, the same $1\% \to 100\% \to 10\%$ cosine schedule envelope, $\beta_1 = 0.9$, $\beta_2 = 0.99999$, Hessian initialization 1, clip radius 10^{-3}, and one Monte

Table 7: Optimization settings for the two trained model families.

Field	MAP (AdamW)	IVON
Peak learning rate	1.5×10^{-4}	0.5
Warmup steps	1000	2000
Initial learning-rate scale	0.01	0.01
Final learning-rate scale	0.1	0.1
Schedule	Linear warmup + cosine decay	Linear warmup + cosine decay
Weight decay	1	10^{-6}
Momentum parameters	$(0.9, 0.98)$	$(0.9, 0.99999)$
Stability / curvature init	$\epsilon = 10^{-8}$	Hessian init $= 1$
Gradient control	Norm clip at 1.0	Clip radius 10^{-3}
Training samples per step	Deterministic model	1 Monte Carlo sample
Evaluation samples	Deterministic model	16 (grid/trajectories); 64 (varying $K-1$ or corrupting y_i)

Table 8: Laplace-specific settings.

Field	Laplace approximation
Base solution	Final MAP model
Bayesian scope	Last layer only
Curvature structure	Kronecker-factored Hessian
Prior precision	Equal to MAP weight decay
Posterior samples	16 (grid); 64 (varying $K-1$ or corrupting y_i)
Fit data	Clean in-distribution training data
Role	Post-hoc predictive uncertainty estimation

Carlo weight sample during training. The effective sample size is set automatically to the number of in-distribution training point pairs multiplied by the effective in-distribution task count after augmentation.

The Laplace model is fit post hoc on the final MAP solution. The approximation is restricted to the last layer, uses a Kronecker-factored Hessian, and sets the prior precision equal to the MAP weight decay (no marginal likelihood optimization). The Laplace fit is always performed on clean in-distribution training data, with full constructed contexts of $K = 32$ triplets (31 in-context examples and one query).

E.5 Experiment families

The MAP training grid underlying Last-Layer Laplace sweeps base task count, training-input fraction, and initialization seed. Specifically, it uses $n_{\text{task}} \in \{8, 16, 32, 64, 128\}$, $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, and seeds 0–3. The full Laplace study uses the corresponding final MAP solutions on the same grid.

The primary IVON study uses the same sweep axes over base task count and training-input fraction, with seeds 0–3. Unless otherwise stated, these runs use the default IVON weight decay 10^{-6}, automatic effective sample size resolution, and 100,000 optimization steps, with the same 200,000-step extension for $n_{\text{task}} = 128$.

The aggregation used in the paper is figure-specific. Final-accuracy heatmaps report final-checkpoint accuracy. Uncertainty-decomposition bars take the arithmetic mean across four seeds and show no dispersion intervals.

The all-run analysis includes all $5 \times 6 \times 4 = 120$ IVON runs. A split is stably successful when its final-five median accuracy is at least 0.75; the three regimes are defined by OOD-Train and OOD-Val success under this rule. The event-clock calculation includes all nine stable OOD-Val successes. Its displayed left panel omits the three $(T = 256, \alpha = 0.8)$ runs with seeds 0, 1, and 3, then groups the six retained runs by (T, α) and plots the median time of each event. Its accuracy event is the first raw checkpoint at or above the threshold, and its EU event is the maximum of the centered-three-

checkpoint-median curve. Per-run, per-split EU collapse is $1 - m_{\text{final},5}/m_{\text{peak},3}$, where $m_{\text{final},5}$ is final-five median raw EU and $m_{\text{peak},3}$ is peak smoothed EU. Thus 0 denotes no endpoint decline from the peak and 1 denotes complete decline to zero; the normalization measures proportional decay rather than cross-run calibration. The script `analysis/plot_split_dynamics.py` emits the all-run figure, a run-level derived CSV, and a sensitivity CSV; all timing summaries are retrospective trajectory diagnostics.

The fixed-weight number-of-examples evaluation uses $K = 1, 2, \ldots, 32$, corresponding to zero through 31 completed in-context examples. It uses $n_{\text{task}} = 64$ (effective task count $T = 256$) and $\alpha = 0.8$. Curves show arithmetic seed means with empirical 2.5th–97.5th percentile bands; these are descriptive across-seed ranges, not confidence intervals. IVON uses seeds 0, 1, and 3, Laplace uses seeds 0–3, and both use 64 posterior samples.

The fixed-weight y_i-token-corruption evaluation uses the same task setting and $K = 32$. Replacement probabilities range over $p_{\text{noise}} \in \{0.0, 0.1, \ldots, 1.0\}$. With the standard one-token causal shift, the implemented supervision-position mask selects the preceding y_i input position; consequently, corruption replaces eligible y_i tokens in the completed in-context examples, rather than their z_i answers. The x_i tokens, final query, and evaluation targets remain clean. Curves use the same seed aggregation and posterior-sampling counts as the number-of-examples evaluation.

E.6 Hardware allocation

All runs use one accelerator per experiment. The main training runs use NVIDIA H200/H100/A100 accelerators, while lightweight validation runs use small H200 MIG partitions. The total computational budget used for the experimental campaign is approximately $12{,}000$ accelerator-hours, split roughly equally between development and final runs.