A Bayesian perspective on learning to grok in-context examples

Abdessamed Qchohi[1] **Simone Rossi**[1]

[1]EURECOM, France (`{name}.{surname}@eurecom.fr`)

Abstract

Learning when a model generalizes — rather than memorizes — is a core problem in AI. In-Context Learning (ICL) lets transformers adapt to new tasks at inference time from a handful of examples, yet what drives this generalization remains unclear. We study grokking in ICL through a Bayesian lens, working in a modular arithmetic setting where the model must infer a hidden linear function from in-context examples alone. We show that epistemic uncertainty collapses within the same training regime in which the model groks, providing a label-free signal of the generalization transition that accuracy alone cannot give. We compare two approximate inference methods — IVON and Last-Layer Laplace — and track how their uncertainty estimates evolve across training, task diversity, context length, and context noise. Our results show that, in this modular-arithmetic ICL setting with decoder-only transformers, uncertainty quantification offers a sharp transition-aligned signature of how the model learns.

1. Introduction

In-Context Learning (ICL) has emerged as a powerful paradigm in which Large Language Models (LLMs) adapt to new tasks by conditioning on a few input-output examples in the prompt (Brown et al., 2020), yet the mechanisms that enable this remain poorly understood. Theoretical accounts range from implicit Bayesian inference (Xie et al., 2022; Akyurek et al., 2023) to meta-learning (Von Oswald et al., 2023): pretraining can be interpreted as learning an implicit prior over tasks, while ICL corresponds to approximate inference over latent task variables given the prompt. Bayesian Deep Learning (BDL) provides a natural language for formalizing this view, treating model parameters as random variables and inferring their posterior given the data (Graves, 2011; Blundell et al., 2015). While exact inference is intractable, approximate methods such as Variational Inference (VI) and Laplace approximations make it possible to measure how uncertainty in the parameters evolves during training (Papamarkou et al., 2024). Separately, models trained on structured tasks often exhibit *grokking*: a delayed generalization transition in which test accuracy remains near zero long after training accuracy saturates, then jumps sharply to near-perfect performance (Power et al., 2022). This raises a natural question: does the model's posterior uncertainty reflect this transition, and can it distinguish memorization from learning? In this work, we study what enables grokking in ICL from a Bayesian perspective, tracking how epistemic and aleatoric uncertainty evolve as the model transitions across learning phases. We work in the modular arithmetic setting of He et al. (2024), where each task is a linear function $z = ax + by \bmod p$ to be inferred purely from in-context examples, and apply two complementary approximate inference methods (IVON and Last-Layer Laplace) to quantify uncertainty throughout training. Our contributions are: (1) a Bayesian analysis of grokking in ICL, in a modular arithmetic setting with a decoder-only transformer, showing how uncertainty signals the generalization transition; (2) an empirical demonstration that epistemic uncertainty

exhibits a sharp transition-aligned signature around the point at which the model groks, indicating generalization rather than memorization; and (3) a comparison of approximate inference methods in capturing these learning dynamics. A comprehensive discussion of related work is provided in Appendix A.

2. Experimental Setup

Tasks. We follow He et al. (2024) and train on linear modular arithmetic tasks of the form $z = ax + by \mod p$, where (a, b) is the task vector and (x, y) is the input vector, with all values in $\{0, \ldots, p-1\}$. In our experiments, we set p = 29 as larger values mainly require more pre-training tasks without affecting the phenomena studied here. Modular arithmetic provides a controlled setting for studying grokking, as the tasks have a clear underlying structure that the model can either memorize or learn to generalize. The task vector is never shown to the model, which receives only input-output triples (x, y, z) and must infer the underlying task from the in-context examples. Each equation is tokenized into three numerical. An example of an input sequence is:

$$(x_1, y_1, z_1),\ \ldots,\ (x_{n_{\text{ctx}}}, y_{n_{\text{ctx}}}, z_{n_{\text{ctx}}}),\ (x_t, y_t, \mathbf{?})$$

During training, the model predicts every third token z_i as well as the final target z_t, with the cross-entropy loss computed only on these positions. At evaluation time, we measure accuracy on z_t only, which directly tests whether the model has inferred the correct task from context. We partition task vectors and input examples into seen and unseen subsets — $\mathcal{T}_{\text{id}}$, $\mathcal{T}_{\text{ood}}$ and $\mathcal{X}_{\text{train}}$, $\mathcal{X}_{\text{test}}$, respectively. Crossing these two partitions gives four evaluation splits that correspond to four distinct learning phases: ID Train ($\mathcal{T}_{\text{id}}$, $\mathcal{X}_{\text{train}}$), which measures memorization on seen tasks and seen inputs; ID Val ($\mathcal{T}_{\text{id}}$, $\mathcal{X}_{\text{test}}$), which tests generalization to unseen inputs on seen tasks; OOD Train ($\mathcal{T}_{\text{ood}}$, $\mathcal{X}_{\text{train}}$), which tests whether the model adapts to unseen tasks on seen inputs; and OOD Val ($\mathcal{T}_{\text{ood}}$, $\mathcal{X}_{\text{test}}$), the hardest split, which tests full out-of-distribution generalization to both unseen tasks and unseen inputs. We train exclusively on ID Train and use the remaining three splits to evaluate generalization. This construction lets us track whether the model memorizes, generalizes in-distribution, or generalizes out-of-distribution, and at which point grokking occurs in each regime.

Model architecture. We train a decoder-only Transformer with Rotary Positional Embeddings (RoPE) (Su et al., 2024) with $d = 6$ layers, embedding dimension 512, 4 attention heads, and an MLP widening factor of 4. We tie input embedding and output projection weights and apply weight decay for regularization. Task vectors are sampled via a parallelogram rule, varying one component at a time so the model can build on previously learned structure, and each batch is constructed so that all tasks in $\mathcal{T}_{\text{id}}$ share the same input sequences, limiting inter-task interference.

Approximate Bayesian inference. We evaluate two complementary methods for approximating the posterior over model parameters. Improved Variational Online Newton (IVON) (Shen et al., 2024) integrates uncertainty estimation into training by maintaining a Gaussian variational posterior over all weights, updated via Newton-like steps at a computational cost comparable to Adam. The Last-Layer Laplace Approximation (Daxberger et al., 2021) is a post-hoc alternative: after standard MAP training, it fits a Gaussian to the posterior of the final linear layer using a Kronecker-factored Fisher matrix, leaving all other parameters fixed. Both methods produce a distribution over predictions and are used to track how posterior uncertainty evolves over the course of training. At evaluation time, predictions are obtained by averaging over S weight samples drawn from $q(\boldsymbol{\theta})$.

Uncertainty decomposition. Both approximate inference methods yield a distribution over predictions, enabling a principled decomposition of predictive uncertainty (Rosso et al., 2025). Given S weight samples $\boldsymbol{\theta}^{(s)}$ from the approximate posterior, total uncertainty (TU) is the entropy of the mean prediction; aleatoric uncertainty (AU) is the mean entropy of individual predictions and

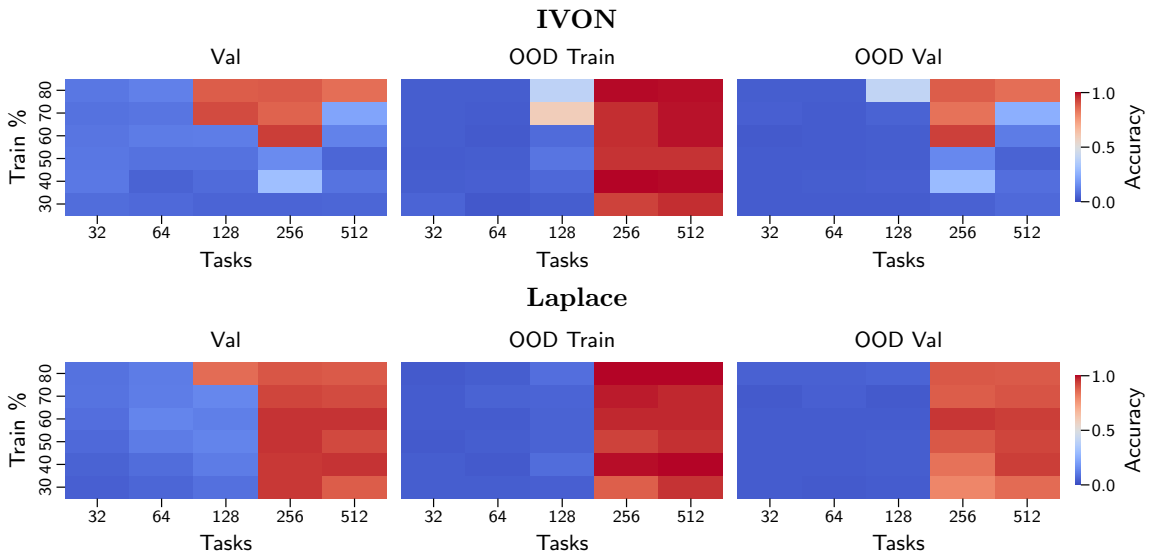

Figure 1: **Accuracy heatmap with IVON and Laplace**. For OOD Train, both IVON and Laplace achieve high accuracy starting at $T = 256$ and $T = 512$ for any training data fraction. For OOD Val (the most challenging split), Laplace generalizes more broadly across configurations, while IVON achieves competitive performance only at high training data fractions.

captures irreducible label noise; and epistemic uncertainty (EU) is their difference:

$$\mathrm{EU}(\boldsymbol{x}) = \mathbb{H}\left[\frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(s)})\right] - \frac{1}{S}\sum_{s=1}^{S} \mathbb{H}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(s)})\right]$$

Epistemic uncertainty reflects the model's lack of knowledge about the task and is expected to decrease as the model acquires more information. In the grokking setting, we use EU as a probe for the generalization transition: it should remain high during the memorization phase and collapse once the model learns the underlying structure of the task.

3. Bayesian In-Context Learning and Grokking

Grokking across task diversity. To study how grokking emerges as a function of task diversity and training data, we train the model across a grid of configurations varying the number of pre-training tasks $T \in \{32, 64, 128, 256, 512\}$ and the fraction of training data $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, running four random seeds per configuration. We measure accuracy on Val, OOD Train, and OOD Val, and compare IVON and Laplace to assess whether the choice of optimizer affects the generalization landscape. Fig. 1 shows the results. Both IVON and Laplace achieve high OOD Train accuracy at $T = 256$ and $T = 512$. For OOD Val, Laplace exhibits stronger and more consistent generalization across task counts and data fractions, whereas IVON reaches competitive OOD Val accuracy only at the highest training data fractions. OOD Val generalization emerges only at the highest task counts for both methods and requires careful tuning of training data fraction.

Epistemic uncertainty tracks grokking. To test whether epistemic uncertainty tracks the grokking transition, we use IVON and Laplace to decompose predictive uncertainty into epistemic and aleatoric components across the same hyperparameter grid. Fig. 2 shows the results in each evaluation set. Both methods show that at low task diversity, total uncertainty is dominated by the epistemic component across all splits. As T increases, epistemic uncertainty drops sharply at $T = 256$ and $T = 512$ for both IVON and Laplace, while aleatoric uncertainty remains high. This

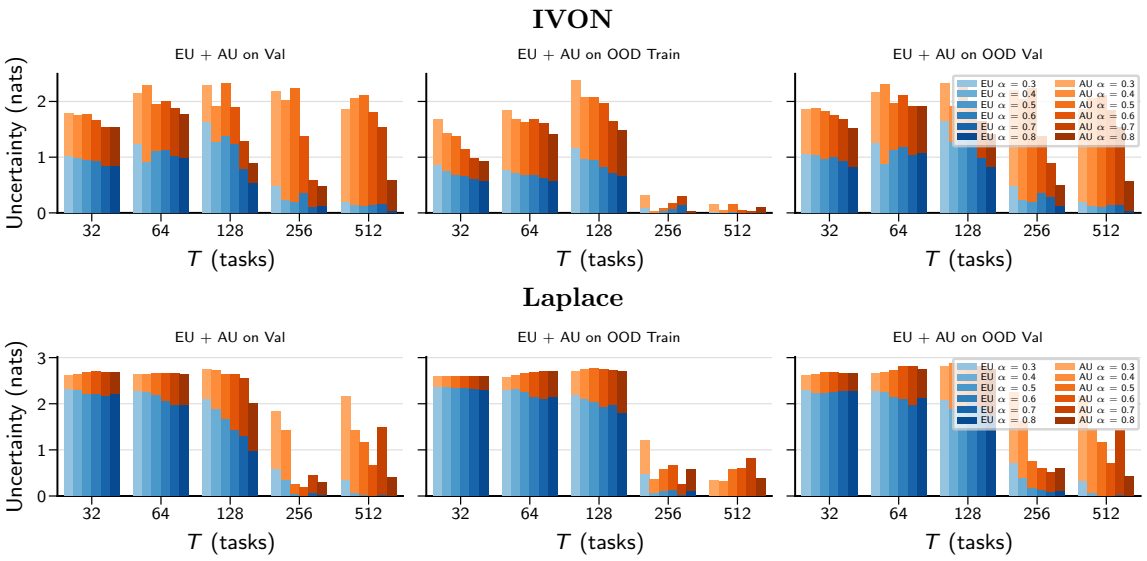

Figure 2: **Uncertainty decomposition with IVON and Laplace**. Decomposition of predictive uncertainty into aleatoric and epistemic components across the pretraining hyperparameter grid.

reduction closely tracks the generalization transition in the accuracy heatmaps, providing a principled uncertainty-based signature of grokking consistent across approximate inference methods. We note that our grid does not include intermediate values between T = 128 and T = 256 (e.g., T = 144, 160), so the exact alignment between the epistemic-uncertainty collapse and the accuracy transition cannot be resolved beyond this granularity. This is an important finding: as of this writing, *grokking* has been primarily characterized in terms of accuracy dynamics (which implies the need for labels to detect it), while here we show that it also has a clear signature in the model's posterior uncertainty, which can be measured without access to labels and provides a more direct window into the model's learning dynamics.

Effect of context length on generalization. To study how the number of in-context examples affects the model's ability to infer the underlying task, we take the models trained with $n_{\text{ctx}} = 32$ and evaluate them by varying the number of in-context examples from 1 to 32 at inference time, without any additional training. The task vector (a, b) is never provided to the model, which must infer it solely from the input-output triples in the context. We compare IVON and Laplace on Val, OOD Train, and OOD Val. Fig. 3 shows accuracy, epistemic uncertainty, and aleatoric uncertainty as a function of context size. Accuracy improves monotonically with context size for both IVON and Laplace across all splits, confirming that more examples indeed help the model infer the task (and thus generalize better on OOD splits). As expected, the aleatoric uncertainty decreases with more context and converges to a floor, reflecting the reduced ambiguity in the task given more evidence. The behavior of epistemic uncertainty is more complex: it increases at small context sizes, likely due to the model's initial uncertainty about the task, then decreases sharply as more examples are added. Notably, this behavior is present in both IVON and Laplace, albeit with different magnitudes, suggesting that it isn't an artifact of the approximate inference method but rather a property of the model's (meta-)learning dynamics as it processes in-context evidence. Also, note that this evidence does not contradict the known fact that uncertainty scales monotonically with data in Bayesian inference (see, e.g., Rosso et al., 2025), but it rather highlights that the current scaling laws of uncertainty should be refined to account for the effect of test-time adaptation via in-context learning.

Robustness to context noise. As a final test of the model's uncertainty estimates, we evaluate how they respond to noise in the in-context examples. For this experiment, we corrupt the context by replacing each label z_i in the context triples with a random token sampled uniformly from

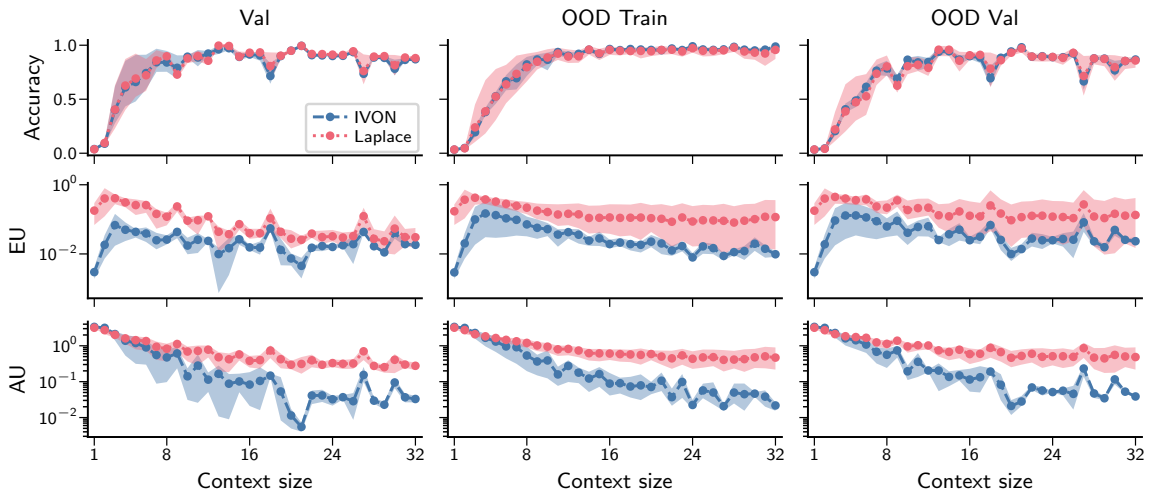

Figure 3: **Accuracy, epistemic and aleatoric uncertainty across context sizes**. Increasing the number of in-context examples monotonically improves accuracy and decreases alepatoric uncertainty, while the epistemic uncertainty follows a non-monotonic behavior.

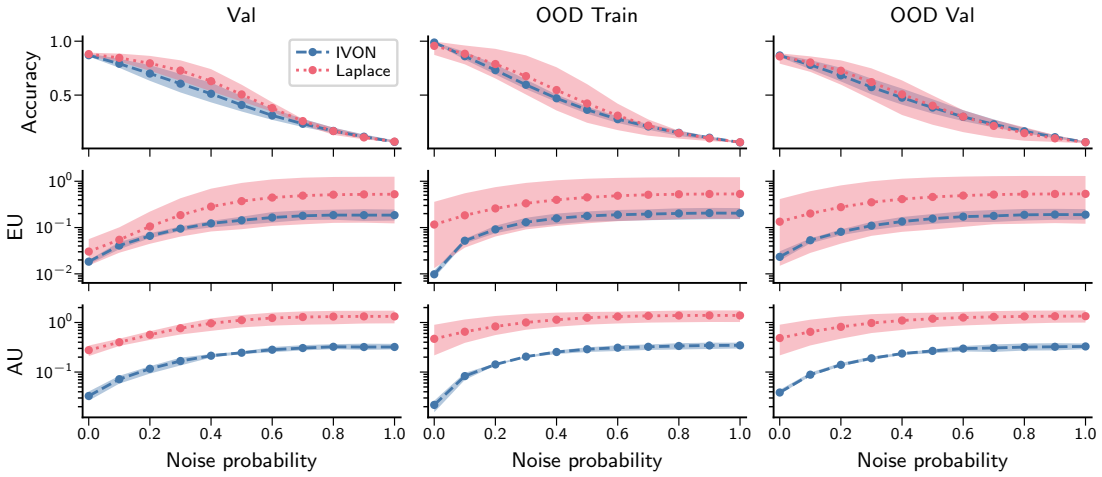

Figure 4: **Accuracy vs. epistemic and aleatoric uncertainty across context noise levels**. Corrupting in-context examples with random label noise degrades accuracy and raises both aleatoric and epistemic uncertainty.

$\{0, \ldots, p-1\}$ with probability $p_{\text{noise}} \in [0, 1]$. The input tokens x and y are never corrupted, and the final query triple is always clean. As we did before, we evaluate both IVON and Laplace on Val, OOD Train, and OOD Val, measuring accuracy, aleatoric uncertainty, and epistemic uncertainty as a function of the noise level (Fig. 4). Accuracy degrades smoothly as noise increases for both methods and all splits. Aleatoric uncertainty rises proportionally to the noise level, correctly capturing the increased label ambiguity in the context. Epistemic uncertainty also increases with noise, but more markedly for Laplace than for IVON, suggesting that the variational posterior learned by IVON is more robust to corrupted context. This is slightly surprising, as one might expect that corrupted context would primarily increase epistemic uncertainty, since it makes the task harder to infer, while aleatoric uncertainty should remain stable as it reflects irreducible noise in the data rather than lack of context. This demonstrates that recent analysis on disentangling aleatoric and epistemic uncertainty (e.g., Mucsányi et al., 2024; Kirchhof et al., 2025) do not necessarily apply to the in-context learning setting, and that further work is needed to understand how these uncertainty components interact when the model is performing test-time adaptation.

4. Concluding Remarks

We studied grokking in in-context learning from a Bayesian perspective, analyzing how uncertainty evolves as transformers trained on modular arithmetic tasks transition from memorization to generalization. Using IVON and Laplace Approximation, we empirically showed that epistemic uncertainty exhibits a sharp transition-aligned signature, collapsing within the same training regime in which the model generalizes, while aleatoric uncertainty remains stable. We further demonstrated that in-context examples act as Bayesian evidence, progressively reducing epistemic uncertainty as context size grows towards the number of examples seen during training, and that IVON's variational posterior is more robust to corrupted context than Laplace. Together, these results suggest that uncertainty quantification provides a principled method for understanding grokking in ICL, and that Bayesian methods offer a promising framework for studying the learning dynamics of transformers beyond simple accuracy metrics.

References

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.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/pdf?id=aVh9KRZdRk.

Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. Position: Uncertainty quantification needs reassessment for large language model agents. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=Lrv20S5RZV.

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.

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, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 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.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in neural information processing systems*, 37:50972–51038, 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. *arXiv*, 2022.

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, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 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*, pages 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, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 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/.

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.

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, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 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. Related Work

Grokking and 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; **?**). 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 (Graves, 2011; Blundell et al., 2015; Gal and 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 an extensive phase-diagram evidence and identified multiple learning phases, emphasizing structured representation emergence as a driver of generalization. **?** focused 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 leaving 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. Notably, we identify the condition under which the model can generalize in-domain and out-of-domain, and we analyze how the model's uncertainty evolves as it transitions from overfitting to generalization, providing a novel perspective on the grokking phenomenon.

B. Methodology

Laplace Approximation. The Laplace Approximation (LA) (Daxberger et al., 2021) is a post-hoc method that constructs a Gaussian approximation to the posterior centered at a Maximum A Posteriori (MAP) estimate θ_{MAP}, obtained by minimizing the regularized empirical risk:

$$\theta_{\text{MAP}} = \arg\min_{\theta} \; -\log p(\theta) - \sum_{n=1}^{N} \log p(y_n \mid x_n, \theta)$$

The approximation follows from a second-order Taylor expansion of the negative log-posterior around θ_{MAP}:

$$p(\theta \mid \mathcal{D}) \approx \mathcal{N}(\theta \mid \theta_{\text{MAP}}, H^{-1})$$

where $H = \nabla_\theta^2 \mathcal{L}(\mathcal{D}; \theta)|_{\theta_{\text{MAP}}}$ is the Hessian of the negative log-posterior, encoding the local curvature of the loss landscape around the MAP solution. Since computing and inverting the full Hessian is intractable for large networks, we apply a Last-Layer Laplace Approximation (LLLA), restricting the Gaussian posterior to the final linear layer while keeping all other parameters fixed at their MAP values. The Hessian of this layer is approximated using a Kronecker-factored Fisher matrix $F \approx A \otimes B$, where A captures correlations of layer inputs and B captures correlations of output gradients. The posterior predictive is then obtained by marginalizing over the Gaussian approximation via Monte Carlo sampling.

Improved Variational Online Newton (IVON). Variational inference (Graves, 2011; Blundell et al., 2015) approximates the posterior by optimizing a tractable distribution $q(\theta)$ to maximize the evidence lower bound (ELBO):

$$\mathcal{L}(q) = \mathbb{E}_{q(\theta)}[\log p(\mathcal{D} \mid \theta)] - \text{KL}(q(\theta) \| p(\theta))$$

Shen et al. (2024) optimizes a Gaussian variational distribution $q(\theta) = \mathcal{N}(\theta \mid m, \text{diag}(\sigma)^2)$ directly during training, integrating uncertainty estimation into the optimization process rather than applying it post-hoc. It estimates the Hessian efficiently using the reparameterization trick:

$$\hat{h} \leftarrow \hat{\nabla} \bar{\ell}(\theta) \odot \frac{\theta - m}{\sigma^2}$$

and updates mean and variance via Newton-like steps:

$$m \leftarrow m - \alpha_t \frac{\bar{g} + \delta m}{h + \delta}, \qquad \sigma \leftarrow \frac{1}{\sqrt{\lambda(h + \delta)}}$$

where δ acts as prior precision through weight decay and λ is set to the dataset size to recover the standard ELBO. This yields a computational cost close to the Adam optimizer while maintaining a proper variational posterior over all model weights. At evaluation time, predictions are obtained by averaging over S weight samples drawn from $q(\theta)$.