Token-Level Uncertainty-Aware Objective for Language Model Post-Training

Tingkai Liu   Ari S. Benjamin   Anthony M. Zador
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY, USA
{tiliu,benjami,zador}@cshl.edu

Abstract

In the current work, we connect token-level uncertainty in causal language modeling to two types of training objectives: 1) masked maximum likelihood (MLE), 2) self-distillation. We show that masked MLE is effective in reducing *epistemic* uncertainty, and serve as an effective token-level *automatic curriculum learning* technique. However, masked MLE is prone to overfitting and requires self-distillation regularization to improve or maintain performance on out-of-distribution tasks. We demonstrate significant performance gain via the proposed training objective - combined masked MLE and self-distillation - across multiple architectures (Gemma, LLaMA, Phi) and datasets (Alpaca, ShareGPT, GSM8K), mitigating overfitting while maintaining adaptability during post-training. Our findings suggest that *uncertainty-aware* training provides an effective mechanism for enhancing language model training.

1 Introduction

Training large language models (LLMs) through next-token prediction with a maximum likelihood objective has shown remarkable generalization capabilities in diverse tasks Chung et al. [2022], Ouyang et al. [2022], Touvron et al. [2023a], Wang et al. [2022], Zheng et al. [2023]. The power of the self-supervised training objective lies in its ability to unify a wide range of language modeling tasks (e.g. grammatical correctness, arithmetic, code generation). However, despite their power and wide adoption, LLMs still suffer from issues such as hallucinations that could stem from overfitting to the training data, particularly during the post-training stage, where the size and diversity of the training set are limited.

In the present work, we examine the heterogeneous nature of tokens and their associated aleatoric and epistemic uncertainties more carefully Hüllermeier and Waegeman [2021], Gupta et al. [2024], Fadeeva et al. [2024]. Aleatoric uncertainty refers to the inherent irreducible stochasticity in the data, whereas epistemic uncertainty refers to model limitations that can potentially be reduced with additional information or a better model. We argue that, as opposed to the vanilla maximum likelihood estimation (MLE) objective, further performance gain could be obtained by focusing on learning tokens with high epistemic uncertainties, while avoiding overfitting by maintaining adequate aleatoric uncertainty estimation on remaining tokens. This issue is particularly prevalent in post-training, where the pre-trained model must simultaneously adapt to new response patterns (high epistemic uncertainty tokens) while retaining generalization across diverse tasks.

However, accurate uncertainty estimation requires aggregating predictions over a *large* model ensemble obtained by, for example, stochastic forwards by Monte Carlo Dropout Sampling Gal and Ghahramani [2016] (MCDO). As MCDO could require hundreds of forward passes to converge, a simpler alternative is required to be used during training. We show that the model's predictive loss, which only requires single forward pass to compute, is a good proxy for epistemic un-
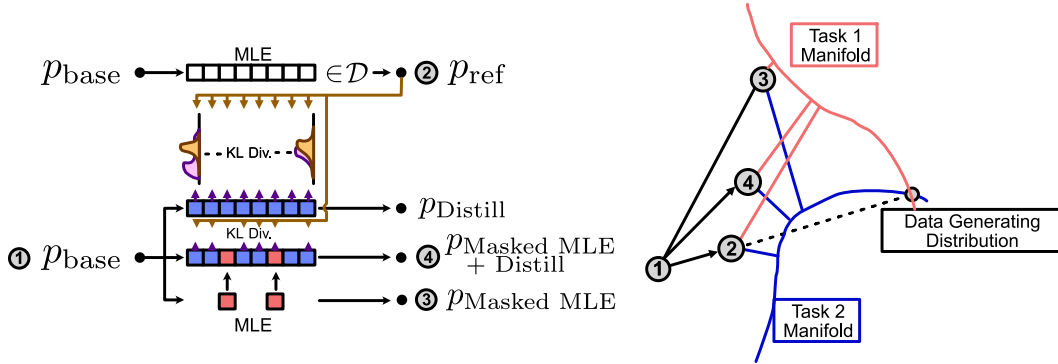
Figure 1: Proposed training procedure combines maximum likelihood with self-distillation training objective to improve both in-distribution and out-of-distribution performances.

certainty estimated via Bayesian Activate Learning by Disagreement (BALD) Houlsby et al. [2011], Kirsch et al. [2019].

Through extensive experiments on multiple architectures (Gemma Team et al. [2024a,b], LLaMATouvron et al. [2023a,b], Grattafiori et al. [2024], PhiAbdin et al. [2024]), datasets (Alpaca, ShareGPT, GSM8K) and downstream tasks (AlpacaEval, IF-Eval, GSM8K), we show that training only on tokens with high loss (masked MLE objective) results in strong in-distribution performance gain compared to vanilla MLE baseline. As epistemic uncertainty reduces with training, the masked MLE objective also provides a natural token-level automatic curriculum learning.

Finally, we observe that training only on tokens with high epistemic uncertainty via the masked MLE objective leads to poor out-of-distribution generalization as a result of overfitting. We show that such issues could be remedied by combining masked MLE on high-loss tokens with a distillation objective on the remaining tokens as shown in Fig. 1.

In conclusion, we propose an uncertainty-aware training objective that outperforms the predominant MLE objective in post-training across both in-distribution and out-of-distribution tasks.

2 Related Works

Training Data Selection Training (core-set) data selection aims to identify a subset of the training data that can effectively represent the entire dataset. Prior arts draw inspiration from a wide range disciplines, ranging from computational geometry Sener and Savarese [2018], gradient-based selection Mirzasoleiman et al. [2020], Xu et al. [2021], Killamsetty et al. [2021], influence functions Wang et al. [2020], statistical mechanics Sorscher et al. [2023] and implicit reward modeling Zhou et al. [2024]. In the current work, we focus on token-level data selection methods Lin et al. [2025].

Curriculum Learning Curriculum learning aims to improve model training by presenting data in a meaningful order, and has been shown to be effective across computer vision Bengio et al. [2009], Weinshall et al. [2018] and language modeling Elman [1993], Xu et al. [2020], Zhang et al. [2018] tasks. In recent years, automated methods of curriculum designs have been proposed based on model competence Platanios et al. [2019], model ranking Sachan and Xing [2016], and gradient norm Liu et al. [2020].

Uncertainty Estimation Uncertainty estimation Hüllermeier and Waegeman [2021] and related areas such as conformal prediction Quach et al. [2024], Yadkori et al. [2024], model calibration Kong et al. [2020], Desai and Durrett [2020] are foundational areas of research especially in the context of large language models Malinin and Gales [2021]. Uncertainty can be estimated via Bayesian model ensembles Pearce et al. [2018] such as Monte Carlo Dropout Gal and Ghahramani [2016] and deep ensembles Lakshminarayanan et al. [2017]. Uncertainty is classified into aleatoric (data/irreducible) and epistemic (model/reducible) types, which can be separately estimated either

via uncertainty estimation heads Nix and Weigend [1994], Kendall and Gal [2017] or Bayesian Active Learning by Disagreement Houlsby et al. [2011].

3 Heterogeneous Token-Level Uncertainty

Traditionally, the training loss of a single datum (i.e. single document in the training corpus) is aggregated across all tokens by taking their average (or sum) of losses. This metric is intuitive in that minimizing such metric is equivalent to maximizing the overall joint likelihood of all tokens in the entire document (we refer to this objective as *vanilla MLE*), which asymptotically converges the model to the data generating distribution (as illustrated in Fig. 1①).
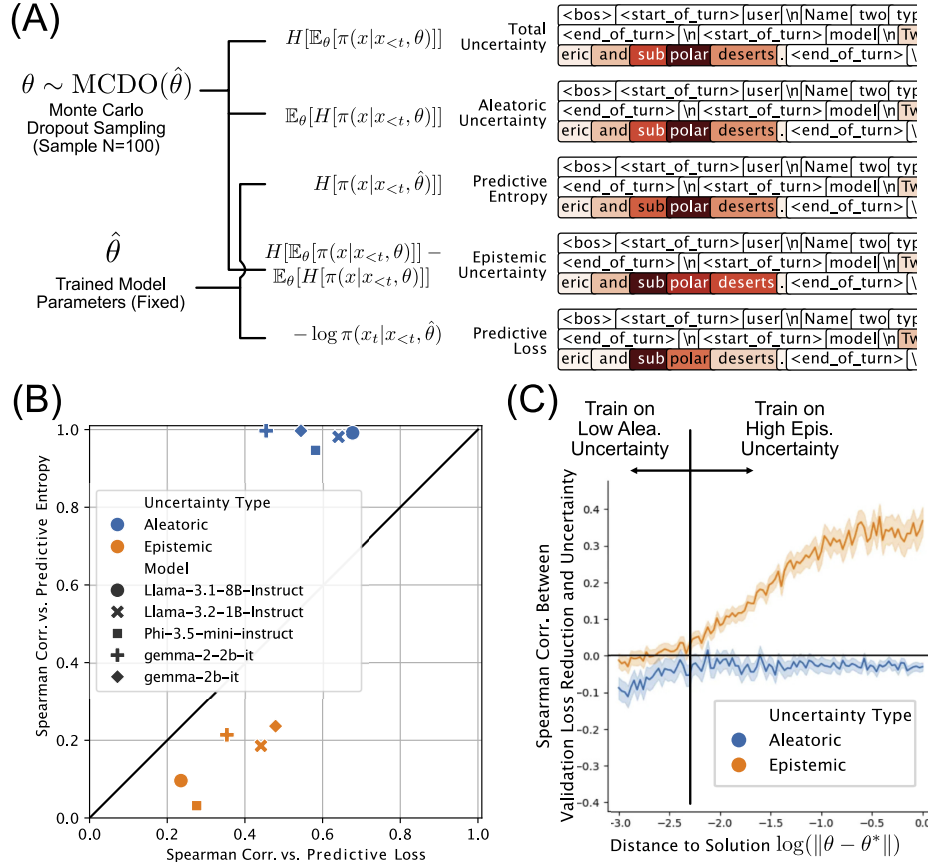

Figure 2: (A) Token level uncertainties, predictive loss and entropy for `Gemma-2B-it`. Note that only tokens in the completion are color coded by the various uncertainty and loss metrics. (B) Correlation between uncertainty (epistemic/aleatoric) and model metrics (predictive loss/entropy) in language modeling of Alpaca dataset across models. (C) Effect of training on different data subset with varying degree of aleatoric/epistemic uncertainty varied based on distance from solution.

However, equally weighting all tokens is not the optimal objective due to (at least) two reasons. First, when data has input-dependent noise that is not uniform across all tokens, this will affect the likelihood in a per-token and heterogeneous manner. This is a case of *aleatoric* noise, referring to noise inherent to the data and due to the environment. In this case it is proper to weigh low-noise tokens higher according to the aleatoric noise level.

A second, competing force is the effect of reducible model uncertainty, called *epistemic* uncertainty, which decreases with model training. If this is known, a Bayesian active learning framework prescribes that the most important data points (tokens) are those which maximally reduce model uncertainty. These can be intuitively understood as the most surprising examples (tokens).

It is important to recognize that the appropriate weights in the MLE objective for each token differ in the aleatoric and epistemic components, and that this tradeoff changes with model training. To demonstrate this in a well-understandable example, we explored a toy linear regression model where we artificially inject known heteroscedastic aleatoric noise, and compared reduction in validation loss by training on single datum. We show that, in linear regression, training should focus on reducing epistemic uncertainty when far from optimum and focus on avoiding data with high aleatoric uncertainty when close to optimum (see Fig. 2(C) for results and Appendix. C.1.1 for detailed description of the experimental procedure). This observation is confirmed by a toy MNIST classification example as shown in Appendix C. In the beginning of training, the epistemic uncertainty is uniform across examples, and it is preferable to prioritize low-aleatoric-noise examples. Later, the epistemic uncertainty decreases faster for low-aleatoric-noise examples, and the important points become the high-epistemic-uncertainty examples.

One of our central claims is that these effects are exaggerated in LLM training due to the nature of the task. LLM training is inherently a massively multi-task endeavor at the token level: individual tokens contribute to a wide range of linguistic and cognitive tasks. This multi-task nature of language model is made apparent when comparing the uncertainty levels at individual tokens. We observe in Fig. 2(A,B) that, across models and datasets, epistemic (resp. aleatoric) uncertainty vary greatly between tokens (see also the overall statistics of per-token losses in Fig. 9 in Appendix. A). This high degree of uncertainty variability across tokens can be exploited to bias the training objective depending on whether a token is in high or low epistemic regime.

To construct the aleatoric and epistemic uncertainty levels for LLMs, we rely on Monte Carlo dropout sampling to construct an ensemble Gal and Ghahramani [2016]. Define the output probability of an autoregressive model with parameters θ as $\pi(x|x_{<t}, \theta)$. We can estimate the aleatoric uncertainty as the entropy of the output classes, marginalized over the ensemble:

$$U_{\text{aleatoric}} = \mathbb{E}_\theta[H[\pi(x|x_{<t})]]$$

The epistemic uncertainty can be calculated via Bayesian Active Learning by Disagreement Nix and Weigend [1994], Kendall and Gal [2017].

$$U_{\text{epistemic}} = H[\mathbb{E}_\theta[\pi(x|x_{<t})]] - \mathbb{E}_\theta[H[\pi(x|x_{<t})]]$$

While these metrics provide useful understanding about how and when tokens should be weighted, they are not practical algorithms for training as they require multiple (hundreds) forward passes to estimate. Here, we report that epistemic and aleatoric losses have definite correlations with the model's predictive loss and output entropy as shown in Fig. 2(B). In particular, the epistemic uncertainty is more correlated with predictive loss than entropy, vice versa for aleatoric uncertainty. This means that useful approximations can be used in practice to achieve higher performance gain.

Given these results, we propose to aggregate MLE loss only on tokens with high loss during training (masked MLE). Simultaneously, we can avoid overfitting by incorporating self-distillation which preserves information regarding aleatoric uncertainty of the remaining tokens.

4 Experiments

4.1 Experimental Setup

Given resource constraints, for our experiments on finetuning pretrained LLM using different masked objectives, we chose three smaller base models Gemma-2B, Gemma-2-2B, and Llama-3.2-1B. All models were trained using rank-32 Low Rank Adaptation (LoRA) on all linear modules with $\alpha = 64$. Unless specified otherwise, all models were finetuned for 1 epoch on the training dataset with batch-size 32, at learning rate `1e-4` with cosine learning rate schedule.

Training & Evaluation Datasets Models were trained and evaluated using a wide range of datasets/tasks as shown in Table 1. Each trained model is evaluated on all tasks to gauge both in-distribution and out-of-distribution performances.

The performance evaluations presented in the current work focus on comparison of training with modified objectives against the *baseline* (vanilla MLE). For freeform QA tasks such as AlpacaEval,

4

Task	Training	Evaluation	Train/Test Size
Single-Turn QA	Alpaca	AlpacaEval 2.0	52K/805
Multi-Turn QA	ShareGPT	-	52K/-
Instruction Following	-	IF-Eval	-/541
Math. Reasoning	GSM8K	GSM8K	7.5K/1.5K

Table 1: Training and Evaluation benchmarks for each task considered in the current work. Note that model trained using each training dataset is evaluated against all downstream tasks to gauge both in-distribution and out-of-distribution performances.
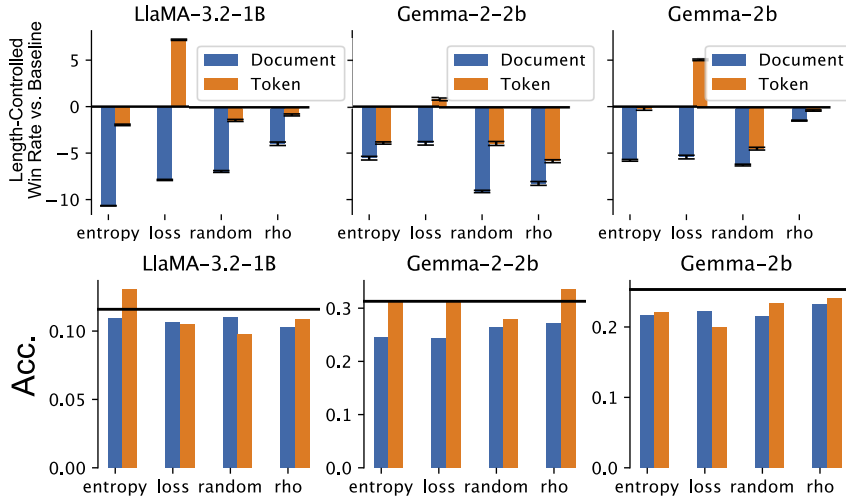

Figure 3: In-Distribution performance gain of token-level masked MLE compared to baseline (vanilla MLE) and document-level masked MLE. Models were trained on (top) Alpaca and (bottom) GSM8K via masked MLE objective on tokens wiht top 25% quantile metric value.

the experiment generations (models trained with objectives other than MLE) were evaluated head-to-head to the generations from the baseline (model trained with MLE), adjusted for both length and positional biases. Given the scale of the experiments, we opted to use `Qwen/Qwen2.5-7B-Instruct` model as the judge, which in Oct. 2024 was the highest ranked judge model on Judge Arena that is smaller than 50B in parameter size. For tasks with ground truth metrics such as IF-Eval and GSM8K, we report both the raw performance metrics as well as the normalized the model performance (`(model - baseline) / baseline`).

4.2 In-Distribution Performance of Token-Level Masked MLE

We show that the fine granularity of token-level masked MLE objective provides superior in-distribution performance than document-level loss masking. We choose 25% of document/tokens to compute MLE losses based on either 1) entropy, 2) loss or 3) Reducible Holdout Loss (RHO). For reference, random selection was also included.

We observe that, as shown in Fig. 3 (and Fig. 6(A)), the finer granularity of token-level masked objective offers superior performance to the document-level counterpart. We observe that training with tokens with highest *loss* is the only method that consistently outperformed baseline on AlpacaEval in a statistically significant manner. Henceforth, we refer to *Masked MLE* as training on tokens with highest loss via the MLE objective.

As training progresses, the distribution of epistemic uncertainty naturally progresses from structural patterns (e.g., "<end_of_turn>") to complex content (e.g., arithmetic) as shown in Fig. 4. Thus providing a new approach to automatic curriculum at a token-level, allowing for fine-grained control over the training process.
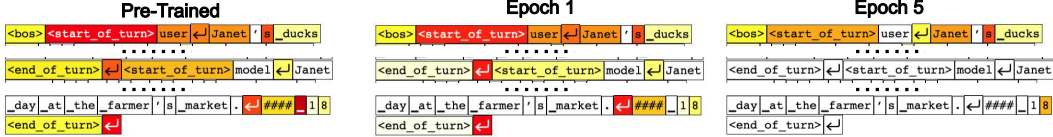
Figure 4: Example of automatic curriculum learning as a result of training on high epistemic uncertainty (color-coded) tokens. Note that while both prompt and response tokens are color coded, losses are only computed and propagated for response tokens during training.
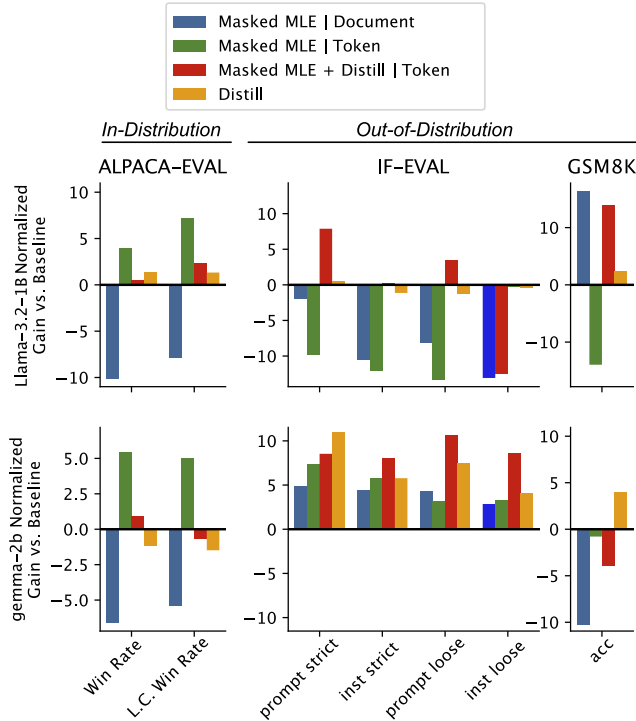

Figure 5: Downstream performance of (top) Llama-3.2-1B and (bottom) Gemma-2B trained on Alpaca dataset with different training objectives.

4.3 Regularization via Self-Distillation Improves OOD Performance

The improved performance of masked MLE comes at the cost of overfitting to the training dataset. As shown in Fig. 5 and Fig. 6(A), the improved Alpaca-Eval performances from finetuning Llama-3.2-1B and Gemma-2B on Alpaca dataset via masked MLE objective result in deterioration of both IF-Eval and other downstream task performances.

To address the issue of over-fitting, we propose to incorporate self-distillation to ensure aleatoric uncertainty is captured by the model on tokens with low epistemic uncertainty. Thus resulting in the final training objective of *Masked MLE + Distill* as shown in Fig. 1 and Fig. 6, given as:

$$\mathcal{L}_t^i = \begin{cases} -\log p_\theta(x_t^i | x^i < t) & \text{high loss} \\ \sum_{x \in \mathcal{V}} -p_{\text{ref}}(x | x_{<t}^i) \log p_\theta(x | x^i < t) & \text{otherwise,} \end{cases}$$

where p_{ref} is the base model finetuned on the same dataset via the vanilla MLE objective.

As shown in Fig. 5, this training objective simultaneously improves in-distribution performance and out-of-distribution generalization for both Llama-3.1-1B and Gemma-2B model trained on Alpaca and ShareGPT (see Fig. 6 (right)) datasets.

We note that the 25% quantile was chosen arbitrarily and tuning of this hyperparameter could potentially lead to stronger overall performance, which we leave for future studies.
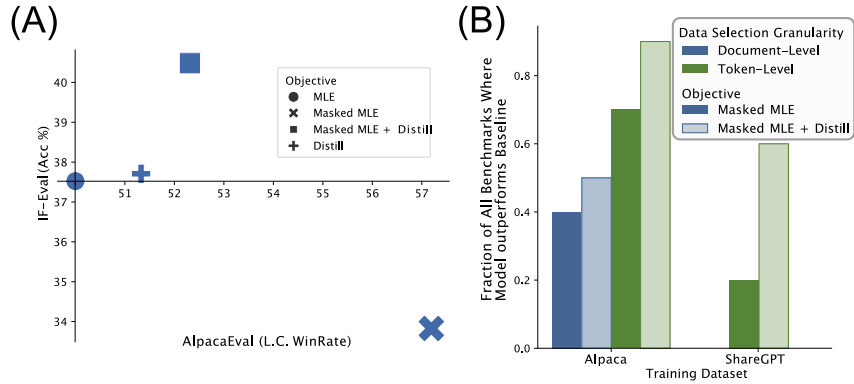
Figure 6: (A) Single-Turn QA (in-distribution) and Instruction-Following (OOD) performance gain over MLE baseline. (B) Fraction of all tasks where a training objective beating baseline.

5 Limitations

Data Selection Granularity and Computational Savings While token-level selection provides fine-grained control over training, its computational benefits are more limited compared to document-level selection due to the underlying mechanics of transformer computation. Token-level selection requires the full model forward and backward, with savings only in the final classification layer which is negligible for very large models.

This modest computational benefit of token-level reselection suggests its primary value lies in its ability to induce better learning dynamics and approximate reward signals, rather than in training efficiency. Future architectures that allow for more efficient sparse attention computation could potentially improve these savings, but with current transformer implementations, document-level selection remains substantially more efficient for reducing computational costs.

Data Curriculum and Model Curriculum While our token-level selection method demonstrates the emergence of an effective data curriculum, the model capacity does not change commensurately. A more comprehensive curriculum would jointly optimize both the data distribution and model complexity, and we leave the exploration of the Pareto frontier of data and model curricula for future works.

6 Conclusion

In conclusion, we show that uncertainty estimation provides a new way of examining training objective in language model at a token-level. We propose a novel training objective that combines masked maximum likelihood and distillation objective that improve model performance on in-distribution and out-of-distribution downstream tasks.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karam-patziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mi-tra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Ros-set, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xi-aoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 295–302, 2020.

Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification, 2024. URL https://arxiv.org/abs/2403.04696.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Ko-renev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind That-tai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Kore-vaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay

Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong,

Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond, 2024. URL https://arxiv.org/abs/2404.10136.

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*, 110(3): 457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3. URL https://doi.org/10.1007/s10994-021-05946-3.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, pages 5574–5584, 2017.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Gradmatch: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474, 2021.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Lingkai Kong, Zhuohan Yao, and Haotian Li. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1326–1340, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need, 2025. URL https://arxiv.org/abs/2404.07965.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems*, 2020.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020.

D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, 1994. doi: 10.1109/ICNN.1994.374138.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International Conference on Machine Learning*, pages 4075–4084, 2018.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1162–1172, 2019.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling, 2024. URL `https://arxiv.org/abs/2306.10193`.

Mrinmaya Sachan and Eric Xing. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 453–463, 2016.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023. URL `https://arxiv.org/abs/2206.14486`.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024a. URL `https://arxiv.org/abs/2403.08295`.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier

Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024b. URL `https://arxiv.org/abs/2408.00118`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:2006.05929*, 2020.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pages 5238–5246, 2018.

Chao Xu, Dacheng Tao, and Chao Xu. On the convergence of learning-based iterative methods for nonconvex inverse problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Gang Xu, Zhiwei Ding, Wenzhan Li, and Yizhou Dong. Gradient-driven rewards to guarantee fairness in collaborative machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention, 2024. URL https://arxiv.org/abs/2405.01563.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Matt Post. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*, 2018.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Haotian Zhou, Tingkai Liu, Qianli Ma, Yufeng Zhang, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. Davir: Data selection via implicit reward for large language models, 2024. URL https://arxiv.org/abs/2310.13008.
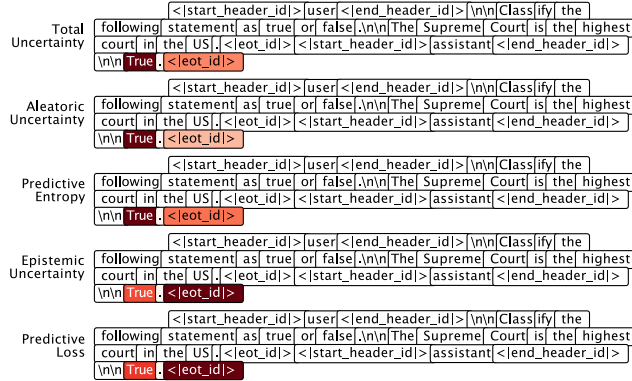
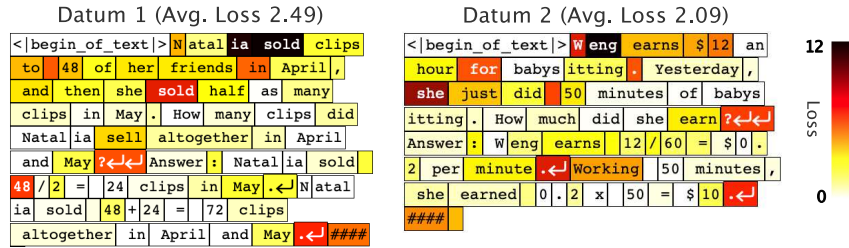Figure 7: Token level uncertainty and loss metrics for `Llama-3.2-8B-Instruct`.

Figure 8: Token-level and datum-level losses of Phi-3.5-Mini-Instruct on two data points in the GSM8K dataset.

A Token-Level Uncertainty via Dropout Ensemble

LLM training is inherently a multi-task as individual tokens contribute to a wide range of linguistic and cognitive tasks, such as:

- Sentiment analysis (e.g., "The movie was $\boxed{excellent}$!")
- Grammar and syntax (e.g., "The cat sat $\boxed{on}$ the mat.")
- Variable definition in code generation (e.g., "let $\boxed{x}$ = 5;")
- Arithmetic computation (e.g., "7 * 6 = $\boxed{42}$")

Which is demonstrated by both the high degree of loss variability at the token-level as shown in Fig. 9, as well as the example in Fig. 7.

Here token-level uncertainties are estimated by Monte Carlo Dropout Sampling (MCDO) with 100 samples at dropout rate of 0.1. The total uncertainty is given by the entropy of the empirically averaged predictive distribution, aleatoric uncertainty is given by the empirical average of entropy of each MCDO sampled predictive distribution and epistemic uncertainty is their difference. In effect, the epistemic uncertainty is equivalent to the Bayesian Active Learning by Disagreement (BALD) objective. We note that measures of epistemic uncertainty should increase monotically with parameter noise (dropout rate), which is confirmed for the BALD objective as shown in Fig. 10.

B Weighted Classification in Sequential Generative Modeling

In this section we formulate the weighted classification objective which unifies a wide range of training objectives as well as data selection methods in sequential modeling.

Consider the following objective

$$\max_{\theta} \sum_{(i,t)} \sum_{x \in \mathcal{V}} w_{i,t}(x) \; \log p_{\theta}\big(x | x_{<t}^i\big) \tag{1}$$
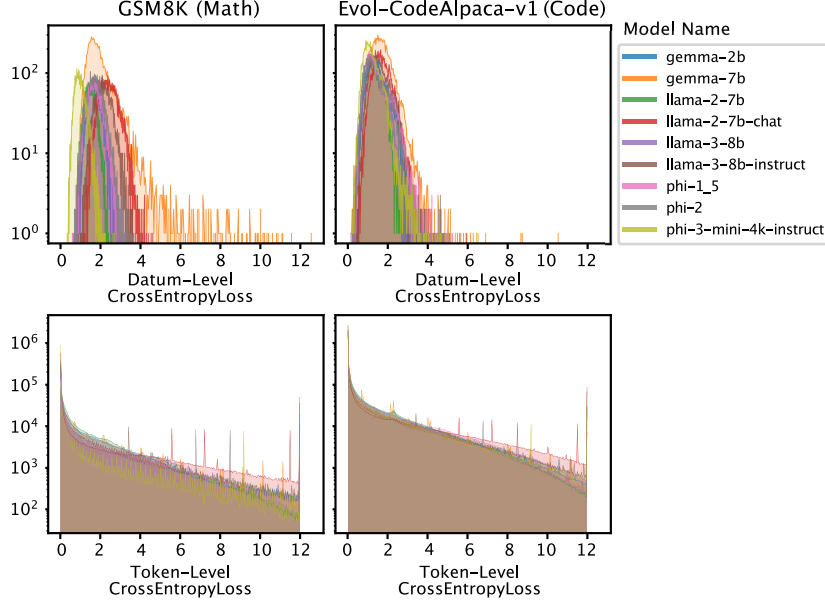
14

Figure 9: Distribution of datum-level and token-level losses across models and datasets.
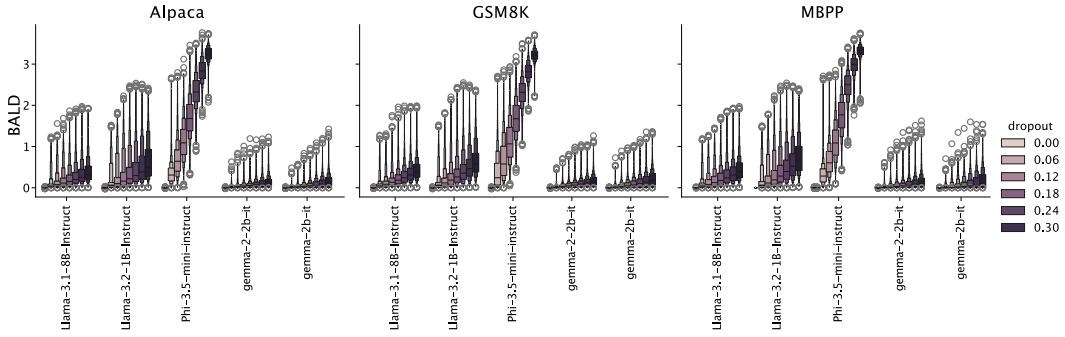

Figure 10: Distribution of BALD metric across dropout rate for different models. Note that BALD scales monotonically with increasing parameter noise, consistent with the notion of epistemic uncertainty. Also note that different models have different degree of sensitivity to dropout ratio, partly due to the difference in the number of dropout modules in the model. For example `Phi-3.5-mini-instruct` have 3 dropout modules at various levels of the model whereas LlaMA models only have attention dropout.

where $x^i_{<t}$ is the context (the tokens before position t in sequence i), $x \in \mathcal{V}$ ranges over the *vocabulary* $\mathcal{V}$ of size $|\mathcal{V}| = V$ and $w_{i,t}(x)$ is a token-level *weight* that scales the log-likelihood of predicting token x given context $x^i_{<t}$.

The formulation in (1) is general by design, admits a wide range of training objectives as special cases:

- **(Masked) Maximum Log-Likelihood** Equivalent to CrossEntropyLoss with 1-hot labels, the weights are given as
$$w_{i,t}(x) = \delta(x = x^i_t),$$
which is the standard training objective in pre-/post-training (supervised fine-tuning) of language models. This formulation further admits masked MLE objective which encompasses training with core-set selection, where weights are given as
$$w_{i,t}(x) = \delta(x = x^i_t)m^i_t$$
for some token-level mask $m^i_t \in \{0, 1\}$ that either includes or discards a token from the training objective.

15

- **(Forward) Knowledge Distillation** Equivalent to CrossEntropyLoss with full teacher distributions, where the weights are given as.

$$w_{i,t}(x) \;=\; p_{\text{teacher}}(x|x^i_{<t}),$$

This formulation further admits a special case of truncated KL divergence,

$$w_{i,t}(x) \;=\; p_{\text{teacher}}(x|x^i_{<t})\,\delta\big(x \in X^i_t\big),$$

where $X^i_t, |X^i_t| \leq |\mathcal{V}|$ is some set of admissible tokens. Truncated KL divergence has previously been been shown to improve student LM performance performance. Note that the this formulation does not include *reverse* distillation which minimizes the objective $D_{KL}(p_{\text{teacher}}||p_{\text{student}})$ (which includes an entropy term of the student model in the objective). However, since the effectiveness of reverse distillation is unclear, we omit this training objective from our formulation.

- **Policy Gradient** For techniques such as Proximal Policy Optimization (PPO) and Group Relative Policy Optimization (GRPO), the weights are given as

$$w_{i,t}(x) \;=\; A(x|x^i_{<t})$$

where $A(\cdot)$ is the advantage function of token x given context. Note that we omit the KL regularization terms here for brevity.

- **Weighing due to aleatoric (heteroskedastic) noise** When labels are noisy and furthermore this noise on labels is input-dependent, and not uniform across examples, the likelihood function changes. The effective result of this is a per-example weighting of the log-likelihood. In Least-Squares regression, for example, adjusting for the input-dependence variance of noise results in Weighted Least Squares. In a classification setting, one instead considers the *oracle* distribution p_{oracle} as capturing the true aleatoric label noise. If this is known (which is rare, although such datasets are becoming more common, e.g. Wei et al. [2021]), then,

$$w_{i,t}(x) \;=\; p_{\text{oracle}}(x|x^i_{<t}).$$

Note the similarity of this objective to model distillation.

- **Active learning via epistemic uncertainty** In an active learning setting, the goal is to select examples for optimal learning, considering one's current uncertainty. A principled formulation of this is Bayesian active learning, a task-agnostic formulation which aims to decrease the uncertainty over the model parameters. This can be expressed through a data masking function $\delta(x = x^i_t)$ that selects points which maximize expected information gain:

$$\mathbb{E}_{x \sim p_\theta(\cdot|x^i_{<t})}\left[D_{KL}(p(\theta|x^i_{<t},x)|p(\theta|x^i_{<t}))\right],$$

where $p(\theta|x^i_{<t})$ represents the current posterior over model parameters and $p(\theta|x^i_{<t},x)$ is the updated posterior after observing example x. In practice, this expectation can be approximated using ensemble methods or Monte Carlo dropout, leading to computationally tractable uncertainty estimates that guide the selection of informative examples for training (e.g. Kirsch et al. [2019]).

B.1 Comparing Three Special Cases

Assuming we are given training labels $\{x^i_t\}_{i,t}$ sampled from some oracle distribution $p_{teacher}$, and assume that there exists an oracle reward model (and consequently an oracle advantage function A), we argue that the performance of the student model trained via the three above objectives would follow the order of:

$$\text{Policy Gradient} \geq \underbrace{\text{Knowledge Distillation} \geq \text{MLE}}_{\text{Distribution Matching}}$$

Policy Gradient vs. Distribution Matching - Task Alignment It is easy to see that Policy Gradient and distribution matching objectives are *equivalent* when $A(x|x^i_{<t}) = p_{\text{teacher}}(x|x^i_{<t})$. This corresponds to perfect "task alignment", where the teacher model is optimized for the downstream task of interest. This assumption holds true for tasks such as image classification, as evidenced by the high correlation (> 0.95) between model likelihood scores and classification accuracy, and by the asymptotic convergence of maximum likelihood training towards optimal performance (e.g. super-human results on benchmarks like ImageNet).

However, for language modeling, the alignment between downstream task performance and distribution matching to the data generating distribution varies substantially across domains. For example, open domain QA tasks would exhibit higher correlation between likelihood and human preferences as compared to structured reasoning tasks such as maths and code generation.

As shown by DeepSeek-R1, pure reward maximization via policy gradient can lead to state-of-the-art performance on downstream tasks with oracle reward models (e.g. reasoning tasks with objective ground truths). However, for majority of open domain tasks that do not have ground truth reward, training with RL is prone to reward hacking and requires frequent data labeling and retraining of reward model to ensure that the reward model adapts to the changing distribution of the underlying language model. Fortunately, for such tasks, results above shows that distance to data generating distribution is predictive of task performance.

C Data and Model Curriculum in Toy Models

In this section, we study the effect of biasing training on different samples based on model performance and uncertainty estimations in two toy modes: linear regression and MNIST classification with MLP.

C.1 Linear Regression

Consider the problem of linear regression, where for a linear system of equation $Zw = d$, we want to find the optimal parameter w by solving the following optimization problem:

$$\min_w \mathcal{L}(w) = \min_w \frac{1}{2}\|Zw - d\|_2^2 \tag{2}$$

where for $Z \in \mathbb{R}^{N \times D}$ is the measurement matrix, $w \in \mathbb{R}^D$ is parameters and $d \in \mathbb{R}^N$ is the observation. Note that the linear regression problem can be applied to a wide range of problems including both polynomial regression and sequence modeling (e.g. autoregressive modeling).

The linear optimization problem can be solved via gradient descent where the gradient update is given as

$$w^{k+1} = w^k - \lambda \cdot \nabla_w \mathcal{L}(w) = w^k - \lambda \cdot Z^T (Zw - d) \tag{3}$$

Here, we consider Data Subset Selection (Fig. 11a), which is akin to masked MLE objective discussed in the current work. Given a indicator function of the data entry $\mathbb{I}_D \in \{0, 1\}^N$, the updated linear system and the associated gradient update is given as

$$\text{Diag}(\mathbb{I}_D) \cdot Zw = d$$
$$\nabla_w \mathcal{L}(w) \leftarrow (\text{Diag}(\mathbb{I}_D)Z)^T (Zw - d) \tag{4}$$

We formulate the problem of finding the optimal subset $\mathbb{I}_D$ in a greedy fashion: for each training step given current estimate of the parameters w^k, find the optimal subset such that the training loss $\mathcal{L}(w^{k+1})$ is minimized.

For finding the optimal data subset $\mathbb{I}_D$, it can be shown that the optimal $\mathbb{I}_D$ can be written as:

$$\min_{\mathbb{I}_D} \mathcal{L}(w^{k+1}) = \min_{\mathbb{I}_D} \frac{1}{2}\|(I - \lambda ZZ^T \mathbb{I}_D) \underbrace{(Zw^k - y)}_{\epsilon^k}\|_2^2 \tag{5}$$

where w^k is the parameter estimate at training step k, $\epsilon^k = Zw^k - y \in \mathbb{R}^N$ is the corresponding residual error for the training dataset. It is easy to see that the optimal choice of $\mathbb{I}_D$ is dependent both on the current residual error and the covariance of training data ZZ^T. If the training data is uncorrelated (or whiten-ed), which results in the covariance matrix ZZ^T being an identity matrix, then the optimal choice of $\mathbb{I}_D$ is exactly $\text{Diag}(\text{Top-K}(\epsilon^k))$, where the selected data are those that have the largest current residual error.

While the results above are only exact if very strong assumptions are place on the data matrix Z, we found that, in practice, these heuristics remain highly effective for arbitrary Z. In Fig. 11, we compared the evaluation loss of models trained via the two heuristics on hold-out data against the

(a) Training on data subset

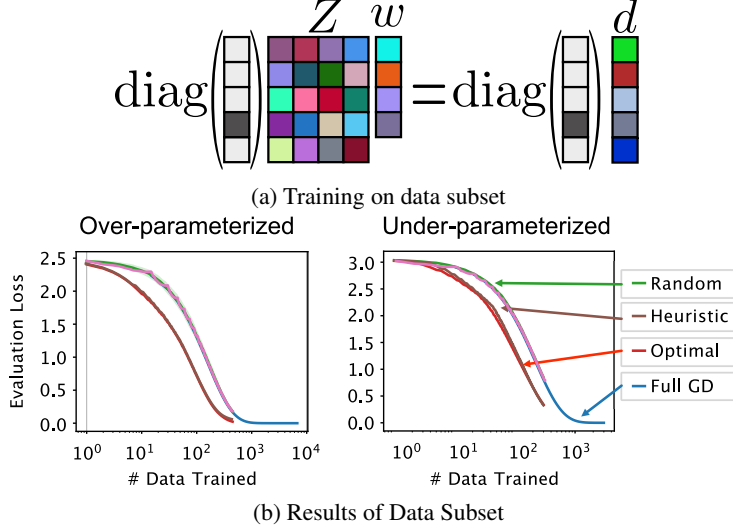
(b) Results of Data Subset

Figure 11: Results of linear regression with training on data subset .

vanilla gradient descent using all of model parameters/gradients and all of the dataset (Full GD). We also performed exhaustive search to find the globally optimal solution of $\mathbb{I}_D$ at each step. Finally, we also include random subset selection as baselines.

As expected, we note that the globally optimal solution (shown in red) that converges to the optimal solution faster than the vanilla gradient descent algorithm for a fixed compute (in terms of number of data trained or number of parameters updated).

In conclusion, we show that data subset selection during training can be done effectively and efficiently via the maximum error heuristics for linear systems.

C.1.1 Effective of Data Uncertainty on Training

To measure the effect of training on data with different degree of epistemic and aleatoric uncertainties, we first created a 5th degree polynomial with random coefficient as ground truth, and sampled 20 (x, y) pairs from the ground truth polynomial as training data. 100 uniformly spaced data in range $x \in [-1, 1]$ and their corresponding y values are computed as held-out validation data.

Aleatoric uncertainty is controlled by adding heteroscedastic Gaussian noise with known standard deviation $(0.1 \cdot \text{Unif}(0, 1))$ to the y value of each of the training datum. Epistemic uncertainty is estimated by first adding 0-meaned Gaussian noise with standard deviation 0.002 to the coefficients of the polynomial, and computing the the variance of 1000 y predictions for each x (with 1000 samples of the Gaussian noise).

We simulate different stages of training by adding 0-meaned Gaussian with standard deviation in range $[10^{-3}, 1]$ to the ground truth coefficients. For each noisy coefficient value, we perform 1 gradient descent step with learning rate 0.01 on each of the 20 data points (with aleatoric noise added), and compute the amount of validation error *decrement* before and after training. This allows us to have accurate per-datum information on the impact of training on a given datum, which we can relate to the amount of aleatoric and epistemic uncertainty of the said datum.

To evaluate how training on data with varying degree of aleatoric/epistemic uncertainty affects model performance at different stages of training, we computed Spearman Ranked correlation between uncertainty level and validation error decrement as shown in Fig. 2(C). We observe that when coefficients are far from ground truth, epistemic uncertainty is important for improving model performance. In contrast, as model converges to ground truth, training on data with high epistemic uncertainty can lead to worsening in model performance, and we instead ought to focus on avoiding overfitting to data with high aleatoric uncertainty.
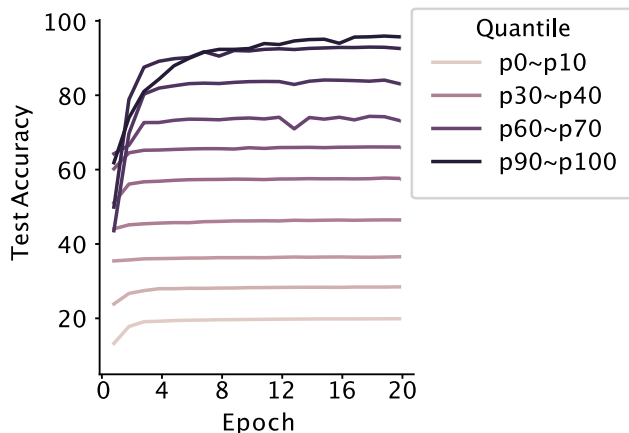
18

Figure 12: Test accuracy vs. training epoch when trained on data in different loss quantile.

C.2 MNIST Classification with MLP

We applied the maximum-error data selection heuristic to the more complex problem of MNIST Classification with MLP model. Note that this problem differs from that of the linear regression problem described in the previous section both in terms of its complexity but also the optimization's convexity, as image classification via MLP is known to be a highly non-convex optimization problem.

We chose a MLP model with two hidden layers of sizes 32 and 16 respectively and `ReLU` activation function. The model is trained on the MNIST training set for 20 epochs using a constant learning rate of 0.01 and batch-size of 256 with Stochastic Gradient Descent (SGD) optimizer.

To validate the effectiveness of the heuristics, we compared the test accuracy of the model trained using data with losses in different quantiles. Note that, as opposed to linear regression, the larger number of training data and parameters makes computing the quantiles for data losses globally (across all training data) infeasible. Instead, we opted to computed the quantiles *in-batch*. In Fig. 12, we show the test accuracy across training epochs for training on 10% subsets of loss. We observe that test accuracy increases almost monotonically with increasing quantiles, suggesting that the heuristics of choosing the data that require the most amount of update remains effective for the non-convex MNIST problem.