Revisiting Warm-Start Training: No Generalization Loss under Standard Training Schemes

Hongjoon Ahn[1]*, **Jinu Hyeon**[2]*, **Hyeonseop Shin**[1]*, **and Taesup Moon**[1,2,3]†
[1] Department of Electrical and Computer Engineering (ECE), Seoul National University,
[2] Interdisciplinary Program in Artificial Intelligence (IPAI), Seoul National University,
[3] ASRI / INMC, Seoul National University
{hong0805, kanasimy, hyeonseop.shin, tsmoon}@snu.ac.kr

Abstract

As large-scale datasets grow, neural networks are increasingly trained in a sequential manner, raising concerns about plasticity loss—a reduced ability to adapt to new data. Prior studies suggest that warm-start training, which continues from previously trained model, yields worse generalization than cold-start training, which reinitializes models at each training phase. However, these works often ignore standard training schemes such as utilizing data augmentation. We revisit this problem under standard training schemes and show, through extensive experiments on various settings, and multiple downstream tasks, that warm-start does not harm generalization compared to cold-start. This finding holds consistently across training from scratch, finetuning of pretranied model, and training of foundation models under a warm-start scenario, indicating that warm-starting is a robust and reliable strategy for large-scale neural network training.

1 Introduction

As the data for training large foundation models grows rapidly, it becomes inevitable to focus on incorporating newly arriving data rather than relying on fixed datasets. A key concern in this setting is the *loss of plasticity* [6, 1, 18, 2, 17, 6, 12, 15]—the gradual reduction in a neural network's ability to adapt to new information as the data distribution evolves in a non-stationary manner. Several works have investigated this issue by attributing loss of plasticity to reduced trainability, namely the diminishing ability of neural networks to fit newly introduced data, typically manifested as persistently high training loss during sequential learning.

However, as emphasized by Berariu et al. [3], the loss of plasticity should be understood not only as reduced trainability—the ability of a network to fit new data—but also as degraded generalization, which remains a central challenge in continual foundation model training. This generalization aspect was first concretely illustrated by Ash and Adams [2], who framed it as the **warm-starting** problem: although restarting from prior weights may stabilize optimization, it often results in poorer generalization. Subsequent works [17, 12, 15] have reinforced this observation, consistently reporting that warm-start training yields higher test errors than cold-start training, where the model is re-initialized from scratch at each phase.

Nevertheless, these studies were largely conducted under restricted conditions that diverge from standard training practices (*e.g.*, omitting widely used techniques such as data augmentation), likely to highlight the observed effects, leaving their conclusions limited in scope. Moreover, little attention

*Equal Contribution.
†Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Continual and Compatible Foundation Model Updates.

has been paid to pretrain–finetune settings that are central to foundation model development, where warm-starting is commonly employed yet its impact remains insufficiently examined.

In this work, we address the warm-starting problem by examining it under more general and practical training settings. We begin by analyzing whether training with a warm-start scenario indeed affects neural network performance when models are trained using a standard training scheme. Here, the standard training scheme for a given architecture is defined as the procedure introduced in its seminal work—for example, training ResNet-18 [9] with the same optimizer, learning rate schedule, and data augmentation strategies as described in the original paper. Adopting this notion of standard training, our extensive experiments demonstrate that training neural networks under a warm-start scenario does not adversely impact the generalization ability of the resulting models. We further extend our investigation to pretrain–finetune settings, which are central to the development of foundation models. Through experiments in these settings, we show that, contrary to prior concerns raised in the literature, warm-starting does not lead to degraded generalization performance compared to cold-starting.

2 Revisiting Warm-Start Training: Conceptual and Empirical Motivation

2.1 Conceptual Background and Motivation

Warm-Start Training Warm-starting refers to the practice of continuing training with the parameters of a network that has already been trained on a smaller dataset. When new data chunks are introduced, the model resumes training from this previously learned state, rather than re-initializing the parameters. The term "warm-start" highlights the idea that the network has been "preheated" through prior training before additional data are incorporated.

Cold-Start Training Cold-starting denotes re-initializing all parameters of the network whenever new data chunks are added. Training thus begins anew each time on the enlarged dataset. The expression "cold-start" emphasizes that, unlike warm-starting, the previously preheated network is reset to a cold state before retraining commences.

Warm-starting bears a close conceptual relationship to *continual learning*, as both paradigms emphasize incremental model updates without reinitializing parameters from scratch. In continual learning, a model adapts to newly arriving data or tasks by leveraging previously learned representations—an idea that closely parallels the warm-start process of resuming training from an already optimized state. Both frameworks share the principle of maintaining continuity in learned weights to accelerate convergence and preserve useful knowledge across training phases.

This connection becomes particularly significant in the era of *foundation models*, which are typically pretrained on large and diverse datasets and later extended through finetuning or incremental adaptation. Such adaptation can be viewed as a form of warm-start training, where the pretrained model provides a preheated initialization for downstream learning. Understanding the dynamics of warm-start training, therefore, offers insights into how foundation models evolve under continual updates—whether through task-specific fine-tuning, domain expansion, or sustained exposure to new data distributions—and whether the reuse of learned parameters inherently constrains or enhances generalization.

Previous studies [2, 22, 12, 17, 15] have demonstrated that warm-starting the training of a neural network upon the arrival of new data leads to inferior generalization performance compared to cold-start training. However, this line of work conducted their analyses under vanilla training, deliberately excluding *standard training schemes* such as using data augmentation, in order to highlight the effects attributable solely to warm-starting [2, 17, 15]. While such experimental settings may reveal the severity of warm-starting in restricted scenarios, they do not provide sufficient evidence to conclude that the problem is pervasive in more general circumstances. Accordingly, in this section, we re-examine whether training a network under conventional training strategies within a warm-start scenario indeed leads to inferior generalization compared to cold-start training.

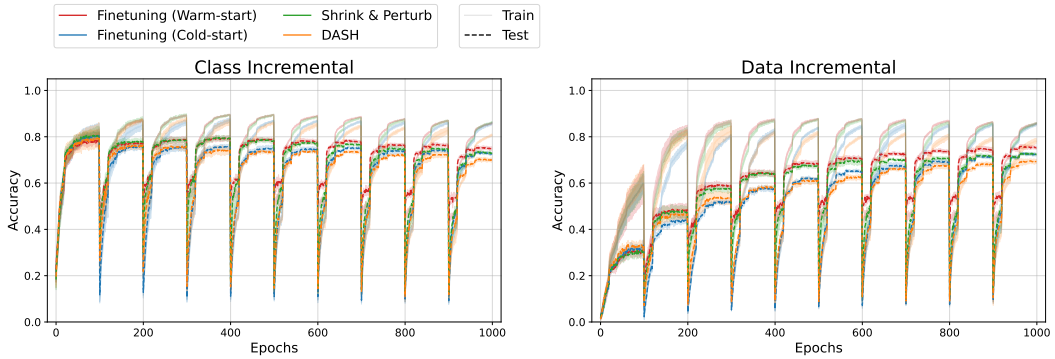

Figure 1: The experiment results on CIFAR 100 dataset for class and data incremental settings. We plot both train and test accuracy.

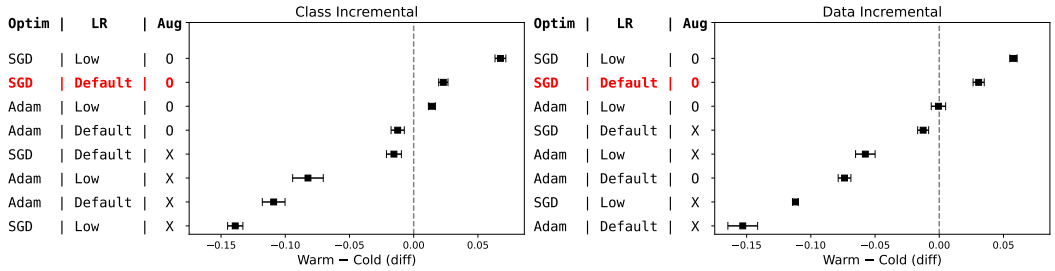

Figure 2: The ablation study on CIFAR 100 dataset with various training schemes. The red colored scheme represents the standard training scheme.

2.2 Empirical Observations on Warm-Start Training

We conduct experiments by training a ResNet-18 [9] model on the CIFAR-100 dataset [11]. The network is optimized using stochastic gradient descent (SGD) with an initial learning rate of 0.1, which is decayed at the 20th and 60th epochs. Standard training practices, including data augmentation, are applied following the original implementation in He et al. [9]. All models are trained for 100 epochs in total.

We evaluate both cold-start and warm-start training settings across two learning scenarios: (1) data-incremental learning, and (2) class-incremental learning. In the data-incremental setting, the original dataset is randomly divided into multiple chunks, while in the class-incremental setting, we first partition the classes randomly and then split the dataset accordingly based on these class subsets. Under this setup, we further compare our results with Shrink & Perturb (S&P) [2] and DASH [17]. All models are trained from scratch for each experiment.

Since our focus is solely on the effect of the warm-start setting, unlike in conventional continual learning, the model has access to all previous data chunks during each training phase. Consequently, we do not employ an external memory buffer containing samples from earlier chunks. This design effectively eliminates the effects of forgetting inherent to continual learning, thereby ensuring that the observed differences arise primarily from the initialization strategy.

Figure 1 shows the results of training the model from scratch. In the figure, the metric reported for both class-incremental and data-incremental settings is **the average accuracy computed over all classes learned up to the current training phase**. This means that the plotted accuracy reflects the model's ability to generalize across all previously learned classes at each point in training. In the class-incremental setting, earlier phases involve learning a smaller number of classes, which makes the initial tasks relatively easier and thus yields higher early-phase accuracy. Conversely, in the data-incremental setting, the model learns all classes from the beginning but with limited data per class in the initial stages, leading to lower early accuracy.

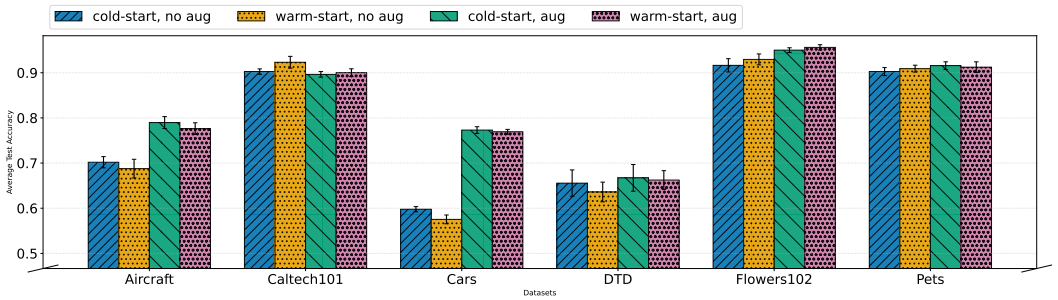

Figure 3: The results of finetuning the pretrained ResNet-18 [9] model on 6 downstream tasks. We report the test accuracies for both warm-start and cold-start, averaged over the entire sequence of training phases.

Different from the phenomena reported in previous works [2, 17, 15], the test accuracy gap between warm-start and cold-start is marginal in the class-incremental setting, and warm-start even performs slightly better in the data-incremental setting. Furthermore, since the generalization ability is not degraded in the warm-start scenario, applying methods such as S&P or DASH offers no advantage. The key takeaway from this experiment is that training under a standard scheme with data augmentation does not yield any meaningful difference in generalization between warm-start and cold-start models. To further analyze this behavior, we conducted an ablation study to identify which components of the standard training setup—namely (1) data augmentation, (2) optimizer, and (3) learning rate—contribute most to generalization.

Figure 2 presents the ablation results, where we visualize and sort the performance difference between warm-start and cold-start across various configurations. The noteworthy finding is that data augmentation emerges as the most influential factor, yet it is often omitted or relegated to the appendices in previous studies [2, 17, 15].

3 Evaluating Warm-Start Training across Finetuning and Pretraining

So far, we have mainly examined the impact of warm-start training when models are trained from scratch, focusing on its effect on generalization ability. However, as large open-source foundation models such as Qwen 2.5 [20], LLaMA 2 [21], and Gemma [19] are widely adopted for solving complex downstream tasks, it becomes crucial to analyze how warm-starting behaves in more realistic scenarios involving pretrained models. To this end, we consider two representative cases where the warm-start setting naturally arises: (1) finetuning a pretrained model on downstream tasks, and (2) pretraining a foundation model under a data-incremental setting. The following subsections present empirical results for both cases.

3.1 Finetuning under the Warm-Start Scenario

In this experiment, we finetune the ImageNet [5] pretrained ResNet-18 [9] model on six downstream datasets: Aircraft [13], Caltech-101 [7], Cars [10], DTD [4], Flowers-102 [14], and Pets [16]. The finetuning follows a data-incremental setting under both warm-start and cold-start scenarios. Each model is trained for 100 epochs per learning phase, and in the cold-start setting, we reload the pretrained model before each new phase. We additionally conduct an ablation study to assess the contribution of data augmentation to generalization performance.

Figure 3 reports the results, where the metric used is the **average test accuracy** computed over the entire sequence of incremental training phases. Consistent with our findings from the training-from-scratch experiments, the performance gap between warm-start and cold-start across all datasets remains marginal. In particular, finetuning under the warm-start setting does not exhibit any degradation in generalization. While the ablation results indicate that data augmentation can still influence generalization in certain cases (e.g., Cars, DTD), its effect is considerably weaker than in scratch training. Overall, these results demonstrate that, regardless of initialization, warm-start finetuning does not inherently impair generalization ability when trained under standard protocols.

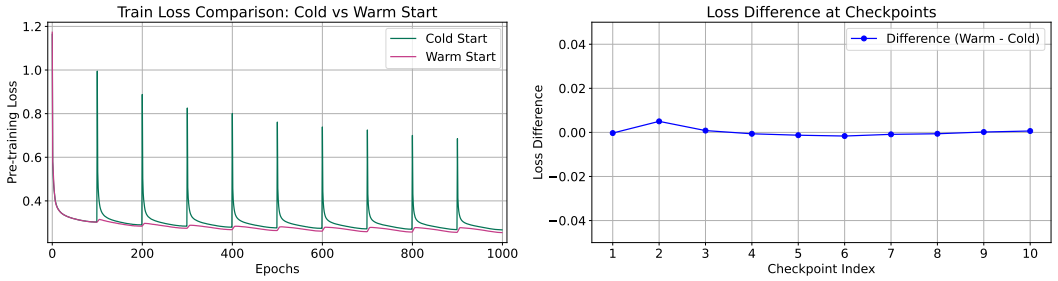

Figure 4: (Left) pretraining loss of both cold-start and warm-start on ImageNet-1K [5] dataset. (Right The pretraining loss differece of loaded checkpoints at each evaluation phase.

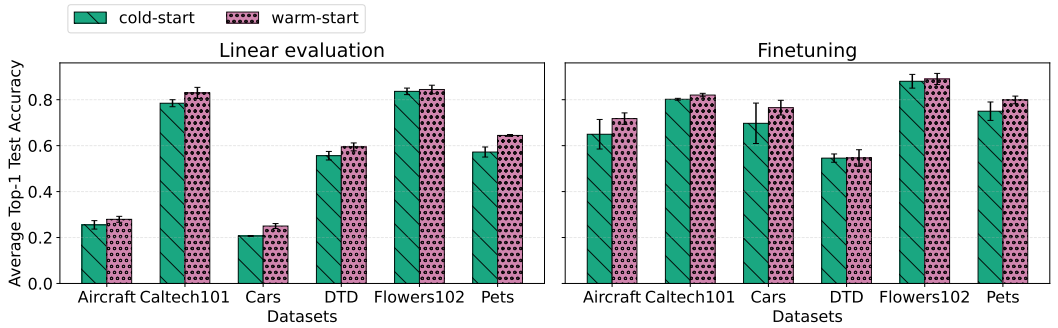

Figure 5: The results of linear evaluation and finetuning on 6 downstream tasks using pretrained MAE. The reported top-1 accuracies represent averages computed across all training phases.

3.2 Pretraining Foundation Models under the Warm-Start Scenario

In this experiment, we pretrain a Masked Autoencoder (MAE) [8] model under a data-incremental setting, applying both cold-start and warm-start strategies. Figure 4 (left) shows the pretraining loss on the ImageNet-1K dataset, where the warm-start consistently achieves lower training loss throughout all phases. To ensure a fair comparison, we align evaluation checkpoints between the two settings such that both models share the same pretraining loss before downstream evaluation.

We evaluate the pretrained checkpoints using both **linear evaluation** and **full finetuning** on the six downstream datasets. The metric reported in Figure 5 is the **average top-1 test accuracy** over all incremental training phases. While minor variations are observed across datasets, the overall performance difference between warm-start and cold-start is negligible for both evaluation protocols. This finding suggests that warm-start pretraining is a stable and reliable strategy, showing no inherent drawbacks in generalization or downstream transfer compared to cold-start initialization.

4 Conclusion and Future Works

In this work, we revisited the warm-start training paradigm through extensive experiments spanning training from scratch, finetuning, and foundation model pretraining. Contrary to prior findings that reported degraded generalization under warm-start settings, our analyses show that when standard training schemes such as data augmentation are properly applied, warm-start training performs on par with or even better than cold-start training. These results suggest that warm-start initialization remains a reliable and efficient approach for incremental and large-scale model training.

This study is limited by its **vision-centric scope** and **empirical nature**. Future work should extend these analyses to other modalities such as language and multimodal models, and develop theoretical insights into why warm-start training maintains generalization under standard optimization schemes.

References

[1] Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C. Machado. Loss of plasticity in continual deep reinforcement learning. In Sarath Chandar, Razvan Pascanu, Hanie Sedghi, and Doina Precup, editors, *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pages 620–636. PMLR, 22–25 Aug 2023.

[2] Jordan Ash and Ryan P Adams. On warm-starting neural network training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3884–3894. Curran Associates, Inc., 2020.

[3] Tudor Berariu, Wojciech Czarnecki, Soham De, Jorg Bornschein, Samuel Smith, Razvan Pascanu, and Claudia Clopath. A study on the plasticity of neural networks, 2023.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 2009.

[6] Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR 2004 Workshop on Generative-Model Based Vision*, 2004.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th IEEE Workshop on 3D Representation and Recognition (3dRR-13), at ICCV*, 2013.

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[12] Hojoon Lee, Hyeonseo Cho, Hyunseung Kim, Donghu Kim, Dugki Min, Jaegul Choo, and Clare Lyle. Slow and steady wins the race: Maintaining plasticity with hare and tortoise networks. In *Forty-first International Conference on Machine Learning*, 2024.

[13] Subhransu Maji, Juho Kannala, Esa Rahtu, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[14] M-E. Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2008.

[15] Sangyeon Park, Isaac Han, Seungwon Oh, and KyungJoong Kim. Activation by interval-wise dropout: A simple way to prevent neural networks from plasticity loss. In *Forty-second International Conference on Machine Learning*, 2025.

[16] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[17] Baekrok Shin, Junsoo Oh, Hanseul Cho, and Chulhee Yun. DASH: Warm-starting neural network training without loss of plasticity under stationarity. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024.

[18] Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32145–32168. PMLR, 23–29 Jul 2023.

[19] 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. *arXiv preprint arXiv:2403.08295*, 2024.

[20] Qwen Team, An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. Technical report, arXiv preprint arXiv:2412.15115, 2024.

[21] 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 Cantón 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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.

[22] Sheheryar Zaidi, Tudor Berariu, Hyunjik Kim, Jorg Bornschein, Claudia Clopath, Yee Whye Teh, and Razvan Pascanu. When does re-initialization work? In Javier Antorán, Arno Blaas, Fan Feng, Sahra Ghalebikesabi, Ian Mason, Melanie F. Pradier, David Rohde, Francisco J. R. Ruiz, and Aaron Schein, editors, *Proceedings on "I Can't Believe It's Not Better! - Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, volume 187 of *Proceedings of Machine Learning Research*, pages 12–26. PMLR, 03 Dec 2023.