Layer Collapse in Diffusion Language Models

Alexander Conzelmann[1,2] A.CONZELMANN@UNI-TUEBINGEN.DE
Albert Catalan-Tatjer[1,2,3] ALBERT.CATALAN-TATJER@TUE.ELLIS.EU
Shiwei Liu[1,2,3] SLIU@TUE.ELLIS.EU
[1]*Tübingen AI Center* [2]*Max Planck Institute for Intelligent Systems* [3]*ELLIS Institute Tübingen*

Abstract

Diffusion language models (DLMs) have emerged as competitive alternatives to autoregressive (AR) language models, yet their activation dynamics remain poorly understood. We characterize these dynamics in LLaDA-8B and identify a striking layer-collapse property: a few early layers exhibit highly similar, collapsed activation patterns dominated by a single large super-outlier that persists across all token positions. Despite its apparent redundancy, pruning this outlier collapses the model into repetitive token loops. Apart from this outlier, LLaDA-8B is more redundant than comparable AR models, with redundancy concentrated in earlier layers, the reverse of the AR pattern, where deeper layers are usually more redundant due to undertraining. Weight spectral analysis attributes this to relative *overtraining* of early layers, and a controlled 160M AR/DLM pre-training pair reproduces the pattern, isolating the diffusion objective as the cause. Code: https://github.com/Conzel/super-outlier-dlm.

Keywords: diffusion language models, layer collapse, layer dynamics

1. Introduction

Diffusion language models (DLMs) have recently emerged as a competitive alternative to autoregressive (AR) large language models. Open models such as LLaDA-8B [31] and DREAM-7B [46] match the quality of AR counterparts at comparable scale on standard reasoning and language understanding benchmarks. Rather than producing one token at a time from left to right, DLMs denoise a masked sequence over multiple refinement steps, enabling parallel token generation [42]. As DLMs become more widespread, questions arise on the structure of their internal representations and how they propagate through layers.

In this paper, we find that LLaDA-8B's internal representations are qualitatively unlike those of any AR model previously studied. Most strikingly, a *single* activation channel (Figure 1) remains persistently and highly activated across all tokens throughout the first half of the model's layers — an extreme outlier that drives layer collapse, causing multiple early layers to produce nearly identical hidden representations. This channel dominates the model to such a degree that ablating it alone causes a near-total collapse in capability, in stark contrast to AR outlier channels [11, 47] whose removal causes only minor degradation. The layer-wise similarity structure is also inverted (Figure 2): standard AR models exhibit distinct early layers and redundant deeper layers [39], while LLaDA-8B is the opposite, with redundancy concentrated in early layers and driven by the super-outlier. Weight spectral analysis attributes this to relative *overtraining* of early layers rather than undertraining.

Contributions.

- We identify a *super-outlier* in LLaDA-8B: a single dominant activation channel whose removal collapses the model.

© A. Conzelmann, A. Catalan-Tatjer & S. Liu.

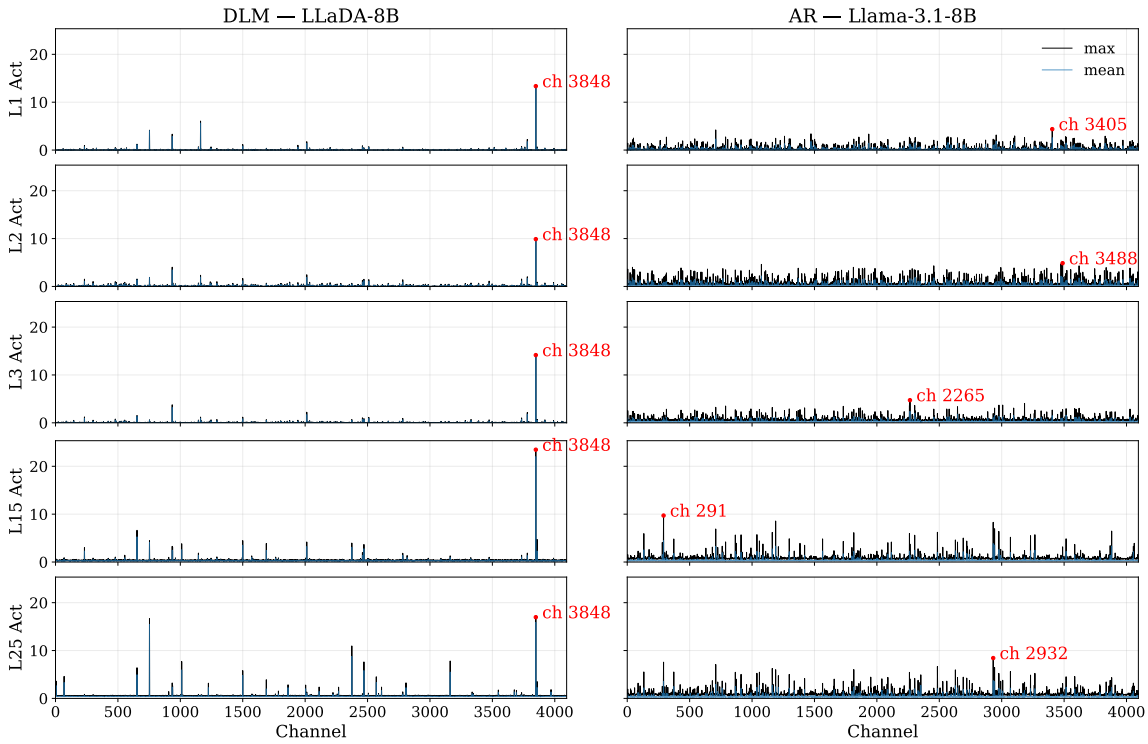


Figure 1: LLaDA activations contain a single consistent super-outlier channel persisting well into the middle layers. The marked dot shows the channel with the largest activation magnitude (averaged over tokens and sequence positions). For Llama, the dominant channel changes on almost every layer.

- We characterize LLaDA-8B's layer-similarity structure and show it is inverted relative to AR models: early layers are highly redundant, with heavy-tailed weight spectra indicating overtraining.
- We pre-train an AR and a DL model under identical conditions to isolate the diffusion objective as the source of these behaviours — to our knowledge, the first such controlled comparison.

2. Layer Dynamics in Diffusion Language Models

2.1. Metrics

Layer similarity. Our main analysis is the per-token cosine similarity between hidden states at different layers. With $h_i(x,t) \in \mathbb{R}^d$ the hidden state at layer i for sequence $x \sim p_x$ at non-padding position t, we define

$$\mathrm{sim}(i,j) = \mathbb{E}_{x,t}\left[\frac{\langle h_i(x,t), h_j(x,t)\rangle}{\|h_i(x,t)\|_2\,\|h_j(x,t)\|_2}\right]; \tag{1}$$

values near 1 indicate near-collinear representations, our proxy for layer redundancy. We estimate activations on 128 sequences of length 2048 from C4; for DLMs we mask a $t \sim \mathrm{Unif}[0,1]$ fraction of tokens, which barely affects the results (subsection C.1).

Weight-spectrum trainability. Similarity alone conflates near-identity redundancy from undertraining with apparent redundancy from a few dominant directions. Following Heavy-Tailed Self-Regularization (HT-SR) theory [26, 27], we also measure each layer's weight-spectrum heavy-tailedness, a data-free quality proxy. For a weight matrix W with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_N$ of $W^\top W$, the Hill estimator on the top k is

$$\alpha_{\text{Hill}}(k) = 1 + \frac{k}{\sum_{i=1}^{k} \log \frac{\lambda_{n-i+1}}{\lambda_{n-k}}}, \tag{2}$$

with smaller α_{Hill} meaning heavier tails (k set by the fix-finger method [45]). A heavier tail indicates stronger feature learning; a balanced α_{Hill} across layers signals a well-trained network, a relatively low α_{Hill} an "overtrained" layer and a high one "undertrained" [17, 23, 51].

2.2. Super-Outlier in LLaDA-8B

Activation outliers in LLMs typically spike only briefly for specific tokens or positions [2]. LLaDA-8B is qualitatively different: a single dominant channel maintains a persistently high activation (up to $5\times$ the next-largest outlier) at *all* sequence positions across many layers (Figure 9). In contrast to AR outliers, which have identifiable mechanisms such as stop-word suppression [47], the super-outlier in LLaDA-8B behaves like a learned constant bias.

Zeroing out the super-outlier channel collapses LLaDA-8B into prompt-independent repetitive token loops (e.g., a GSM8K word problem yields only *"buy buy buy buy1yl buy buy"*): it drops LLaDA-8B from $\approx 83\%$ to 0% on a GSM8K subset, whereas zeroing Llama-3.1-8B's highest-magnitude channel costs only 4 points, so the super-outlier is functionally load-bearing despite the apparent redundancy of the layers it dominates.

2.3. Early-Layer Redundancies in LLaDA-8B

Figure 2 shows pairwise hidden-state similarities. In LLaDA-8B (top left) almost all layers are highly similar, with early layers near-identical over 15+-layer ranges; zeroing the super-outlier (middle) reduces but does not eliminate this. Llama-3.1-8B (bottom left) instead shows distinct early layers and more redundant late ones, as in the AR literature. DREAM-7B (top right) inherits Qwen-2.5-7B's pattern, confirming that diffusion fine-tuning on AR weights does *not* produce the collapsed regime; only training from scratch does, pointing to the training trajectory rather than bidirectional inference or the masking loss as the source of layer collapse.

2.4. Heavy-Tailed Evidence of Overtraining

Weight-spectrum analysis attributes LLaDA-8B's early-layer redundancy to relative *overtraining* rather than undertraining. As shown in Figure 3, early layers in LLaDA-8B have a strikingly low α_{Hill} relative to the rest of the network, while Llama-3.1-8B's distribution is far flatter. The low α_{Hill} co-localises with the super-outlier: an overtrained layer whose representation has collapsed onto a single dominant direction.

3. Supporting Evidence: A Controlled 160M Pair

To rule out architectural or data-mix artifacts, we pre-train two Pythia-160M models [4, 40] on 100B FineWebEdu tokens [32] that differ *only* in the training objective: cross-entropy with masking

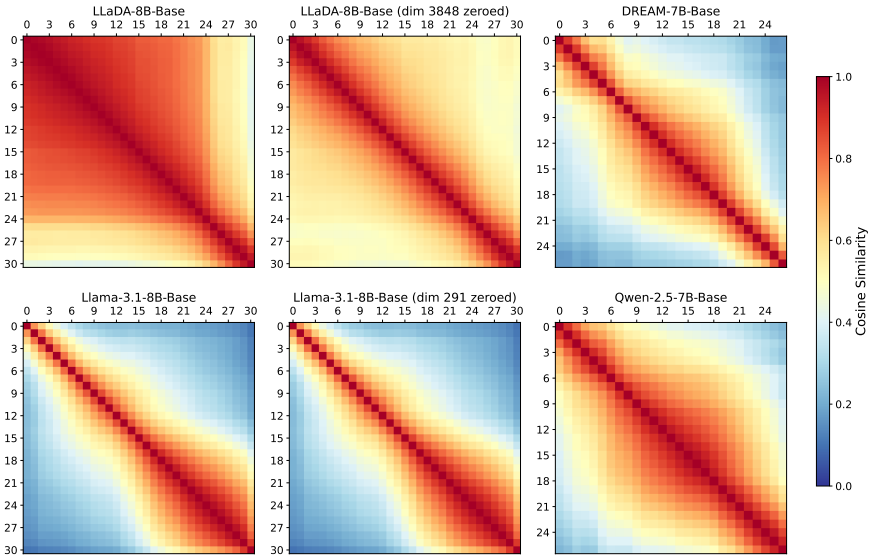


Figure 2: Per-token cosine similarity. Top row: DLMs; bottom row: ARs. Left: LLaDA-8B has highly redundant (collapsed) layer similarities relative to Llama-3.1-8B. Middle: removing the largest outlier from the similarity calculation eliminates much of LLaDA-8B's redundancy but barely affects Llama-3.1-8B. Right: DREAM-7B and Qwen-2.5-7B have nearly identical patterns, as DREAM-7B was fine-tuned from Qwen-2.5-7B weights.

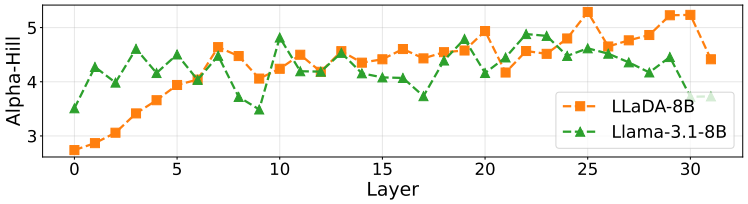


Figure 3: α_{Hill} is very low in the early layers of LLaDA-8B, indicating overtraining; Llama-3.1-8B's distribution is much flatter. Values are averaged over modules within each layer.

(DLM-160M) vs. without (AR-160M); details in subsection B.2. The early-layer redundancy pattern reproduces at this scale (Table 1): block-averaged cosine similarity is higher in DLM-160M than AR-160M within the first 4 layers, and the relation flips within the last 4. The α_{Hill} signature is directionally consistent in the first few layers (lower for DLM-160M). The super-outlier itself does not appear at this scale, consistent with extreme activation outliers only emerging above $\sim$6.7B parameters [11]. The controlled setup isolates the diffusion objective (rather than architecture, data, or bidirectional inference) as the cause.

4. Related Work

Outliers, sinks, and super-weights. A small number of coordinates disproportionately shape LLM behaviour: attention sinks concentrate probability mass on a handful of tokens [16, 43],

Table 1: Average per-token cosine similarity within the first-4 and last-4 layer blocks of the controlled 160M pair (averaged over the 16 inner-block layer pairs). The LLaDA/Llama early-layer-redundancy pattern reproduces.

Model	Early-layer similarity	Deeper-layer similarity
DLM-160M	0.877	0.787
AR-160M	0.818	0.829

massive activations grow orders of magnitude larger than the rest at a few positions [37], and super weights [47] and systematic outliers [2] identify individual parameters and feature dimensions whose removal damages the model. These phenomena are predominantly described along the *token* axis. The super-outlier we report is the *channel-axis* counterpart: a single hidden dimension dominating the representation across all positions, absent in AR models of comparable scale.

Curse of depth. Multiple works analyze the tendency of AR models to have under-trained deeper layers [15, 36, 39], a phenomenon termed the *Curse of Depth*, with mitigations including depth-growing [18], sparsity [29], and mixed Pre/Post-LN [21]. While we show that early layers in DLMs are more redundant than in ARs, this does not mean the curse is broken: even if an improved training paradigm made early layers more distinct, DLMs still exhibit strong redundancies in later layers and could face similar issues.

Weight and activation dynamics in DLMs, and DLM compression. A concurrent work [14] also uses layer cosine similarity and observes early-layer redundancy, but does not identify the super-outlier or perform a controlled AR/DLM comparison; Rulli et al. [33] characterise (token-level) attention sinks in DLMs, complementary to our channel-wide outliers. A separate line of work adapts PTQ and pruning pipelines to DLMs [22, 30, 44, 50]; our focus is analytical, characterising the dynamics rather than proposing a compression method.

5. Discussion

DLMs exhibit activation dynamics qualitatively distinct from their AR counterparts. LLaDA-8B contains a single dominant channel, persistent across all positions in the early layers, whose removal collapses the model, a fragility with no parallel in comparable AR models. Its early layers are also the redundant ones, which we attribute to overtraining (supported by heavy-tailed early-layer spectra) and replicate in a controlled 160M pair. These dynamics also have direct implications for model compression, which we explore in section A.

Limitations. Our large-scale findings rest on a single DLM/AR pair, as LLaDA-8B is, to our knowledge, the only large-scale DLM pre-trained from scratch (other public DLMs [5, 6, 48] are AR-fine-tuned); the 160M pair lacks the super-outlier at that scale.

References

[1] Niccolò Ajroldi. plainLM: Language model pretraining in PyTorch. https://github.com/Niccolo-Ajroldi/plainLM, 2024.

[2] Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[3] Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Straight to zero: Why linearly decaying the learning rate to zero works best for llms, 2025. URL https://arxiv.org/abs/2502.15938.

[4] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 2397–2430. PMLR, 2023. doi: 10.48550/arXiv.2304.01373.

[5] Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, Chengxi Li, Chongxuan Li, Jianguo Li, Zehuan Li, Huabin Liu, Ling Liu, Guoshan Lu, Xiaocheng Lu, Yuxin Ma, Jianfeng Tan, Lanning Wei, Ji-Rong Wen, Yipeng Xing, Xiaolu Zhang, Junbo Zhao, Da Zheng, Jun Zhou, Junlin Zhou, Zhanchao Zhou, Liwang Zhu, and Yihong Zhuang. LLaDA2.0: Scaling up diffusion language models to 100B, 2025.

[6] Tiwei Bie, Maosong Cao, Xiang Cao, Bingsen Chen, Fuyuan Chen, Kun Chen, Lun Du, Daozhuo Feng, Haibo Feng, Mingliang Gong, Zhuocheng Gong, Yanmei Gu, Jian Guan, Kaiyuan Guan, Hongliang He, Zenan Huang, Juyong Jiang, Zhonghui Jiang, Zhenzhong Lan, Chengxi Li, Jianguo Li, Zehuan Li, Huabin Liu, Lin Liu, Guoshan Lu, Yuan Lu, Yuxin Ma, Xingyu Mou, Zhenxuan Pan, Kaida Qiu, Yuji Ren, Jianfeng Tan, Yiding Tian, Zian Wang, Lanning Wei, Tao Wu, Yipeng Xing, Wentao Ye, Liangyu Zha, Tianze Zhang, Xiaolu Zhang, Junbo Zhao, Da Zheng, Hao Zhong, Wanli Zhong, Jun Zhou, Junlin Zhou, Liwang Zhu, Muzhi Zhu, and Yihong Zhuang. LLaDA2.1: Speeding up text diffusion via token editing, 2026.

[7] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language, 2019.

[8] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300.

[9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge, 2018.

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[11] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc., 2022.

[12] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023. URL https://arxiv.org/abs/2210.17323.

[13] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, July 2024.

[14] Raghavv Goel, Risheek Garrepalli, Sudhanshu Agrawal, Chris Lott, Mingu Lee, and Fatih Porikli. A comparative analysis of layer-wise representational capacity in AR and diffusion llms, 2026.

[15] Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Dan Roberts. The unreasonable ineffectiveness of the deeper layers. In *The Thirteenth International Conference on Learning Representations*, 2025.

[16] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Di He, Songjun Tu, Ajay Jaiswal, Li Shen, Ganzhao Yuan, Shiwei Liu, and Lu Yin. AlphaDecay: Module-wise weight decay for heavy-tailed balancing in llms, 2025.

[18] Ferdinand Kapl, Emmanouil Angelis, Tobias Hoppe, Kaitlin Maile, Johannes von Oswald, Nino Scherrer, and Stefan Bauer. Do depth-grown models overcome the curse of depth? An in-depth analysis. *ArXiv*, abs/2512.08819, 2025.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

[20] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[21] Pengxiang Li, Lu Yin, and Shiwei Liu. Mix-LN: Unleashing the power of deeper layers by combining pre-LN and post-LN. *ArXiv*, abs/2412.13795, 2024.

[22] Haokun Lin, Haobo Xu, Yichen Wu, Ziyu Guo, Renrui Zhang, Zhichao Lu, Ying Wei, Qingfu Zhang, and Zhenan Sun. Quantization Meets dLLMs: A Systematic Study of Post-training Quantization for Diffusion LLMs, October 2025.

[23] Zihang Liu, Yuanzhe Hu, Tianyu Pang, Yefan Zhou, Pu Ren, and Yaoqing Yang. Model balancing helps low-data training and fine-tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1311–1331, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.78.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

[25] Aaron Lou, Chenlin Feng, Stefano Ermon, and Jiaming Zhao. Discrete diffusion modeling by estimating the ratios of the transition kernel. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2310.16834.

[26] Haiquan Lu, Yefan Zhou, Shiwei Liu, Zhangyang Wang, Michael W. Mahoney, and Yaoqing Yang. AlphaPruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

[27] Charles H. Martin and Michael W. Mahoney. Traditional and Heavy-Tailed Self Regularization in Neural Network Models, January 2019.

[28] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP*, 2018.

[29] Dilxat Muhtar, Xinyuan Song, Sebastian Pokutta, Max Zimmer, Nico Pelleriti, Thomas Hofmann, and Shiwei Liu. When does sparsity mitigate the curse of depth in llms, 2026.

[30] Aidar Myrzakhan, Tianyi Li, Bowei Guo, Shengkun Tang, and Zhiqiang Shen. Sink-aware pruning for diffusion language models, 2026.

[31] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2026.

[32] Guilherme Penedo, Hynek Kydlíček, Loubna Cappelli, Mario Žilinec, Colin Chapman, Colin Guest, Lucas Guntupalli, Ahmed Bakouch, Igor Malartic, Hugo Touvron, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024. URL https://arxiv.org/abs/2406.17557.

[33] Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. Attention sinks in diffusion language models, 2025.

[34] Subham Sahoo, Aaron Lou, Roger Chen, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.07524.

[35] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale, 2019.

[36] Aleena Siji, Amir Mohammad Karimi-Mamaghan, Ferdinand Kapl, Tobias Höppe, Emmanouil Angelis, Andrea Dittadi, Maurice Brenner, Michael Heinzinger, Karl Henrik Johansson, Kaitlin Maile, Johannes von Oswald, and Stefan Bauer. From words to amino acids: Does the curse of depth persist? *ArXiv*, abs/2602.21750, 2026.

[37] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024.

[38] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2026.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

[41] Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *arXiv preprint arXiv:2410.05192*, 2024.

[42] Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. In *The Fourteenth International Conference on Learning Representations*, 2026.

[43] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Chen Xu and Dawei Yang. DLLMQuant: Quantizing Diffusion-based Large Language Models, August 2025.

[45] Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E. Gonzalez, Kannan Ramchandran, Charles H. Martin, and Michael W. Mahoney. Test accuracy vs. generalization gap: Model selection in NLP without accessing training or testing data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Kdd '23, pages 3011–3021, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 979-8-4007-0103-0. doi: 10.1145/3580305.3599518.

[46] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion Large Language Models, August 2025.

[47] Mengxia Yu, De Wang, Qi Shan, Colorado J Reed, and Alvin Wan. The super weight in large language models, 2025.

[48] Yifan Yu, Yuqing Jian, Junxiong Wang, Zhongzhu Zhou, Donglin Zhuang, Xinyu Fang, Sri Yanamandra, Xiaoxia Wu, Qingyang Wu, Shuaiwen Leon Song, Tri Dao, Ben Athiwaratkun, James Zou, Fan Lai, and Chenfeng Xu. Introspective diffusion language models, 2026.

[49] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence?, 2019.

[50] Tianao Zhang, Zhiteng Li, Xianglong Yan, Haotong Qin, Yong Guo, and Yulun Zhang. Quant-dLLM: Post-training extreme low-bit quantization for diffusion large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[51] Yefan Zhou, Tianyu Pang, Keqin Liu, charles h martin, Michael W. Mahoney, and Yaoqing Yang. Temperature balancing, layer-wise weight analysis, and neural network training. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

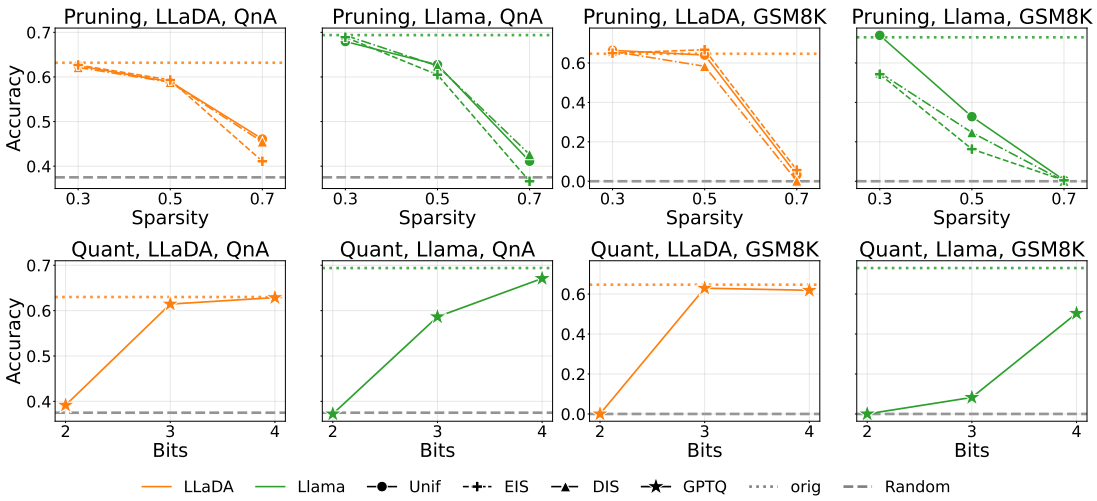


Figure 4: Compression for LLaDA-8B and Llama-3.1-8B: pruning across allocation strategies (top, $\epsilon = 0.08$) and GPTQ across bit-widths (bottom), on base-model QnA (left) and instruct-model GSM8K (right).

Appendix A. DLMs Are More Robust to Compression Despite Super-Outliers

We translate our observations into compression experiments: pruning and quantization of LLaDA-8B and Llama-3.1-8B across overall and layer-wise compression strengths. We find that (a) AR-optimal layer-wise sparsity schedules are *inverted* for DLMs, and (b) DLMs are overall more robust to compression — surprising given the super-outlier's sensitivity, but consistent with LLaDA-8B having fewer secondary outliers and more redundant layers.

Setup. Models are evaluated on the average of 6 QnA tasks (ARC-C, HellaSwag, PIQA, Winogrande, BoolQ, OBQA) and on GSM8K. DLMs use FastDLLM [42] with generation length 1024 and one token per diffusion step. Pruning uses WANDA [38] with 128 C4 samples; quantization uses GPTQ [12] with 256. The schedules *earlier-is-sparser* (EIS) and *deeper-is-sparser* (DIS) are linear with $\epsilon = 0.08$ (layer t in a T-layer model receives sparsity $s \pm 0.08\,(1 - 2(t-1)/(T-1))$). See subsection B.1 for evaluation details.

Pruning. Figure 4 (top row) shows accuracy at sparsities $\{0.3, 0.5, 0.7\}$ under three allocation strategies (uniform, EIS, DIS). Two observations: (i) LLaDA-8B starts below Llama-3.1-8B but is much more robust under compression, achieving almost double the GSM8K accuracy of Llama-3.1-8B at 50% sparsity — consistent with its higher representation redundancy, which makes any single layer less critical (provided the super-outlier is preserved, as WANDA naturally does). (ii) The optimal allocation flips: for Llama-3.1-8B, EIS is always suboptimal and uniform or DIS wins; for LLaDA-8B, DIS is usually the suboptimal choice. Only at 70% sparsity on QnA does DIS beat EIS for LLaDA-8B, likely because that regime forces pruning of weights tied to the super-outlier.

Quantization. The bottom row of Figure 4 shows GPTQ across bit-widths. The same robustness ordering holds: LLaDA-8B starts below Llama-3.1-8B but surpasses it from 3 bits on QnA and 4 bits on GSM8K.

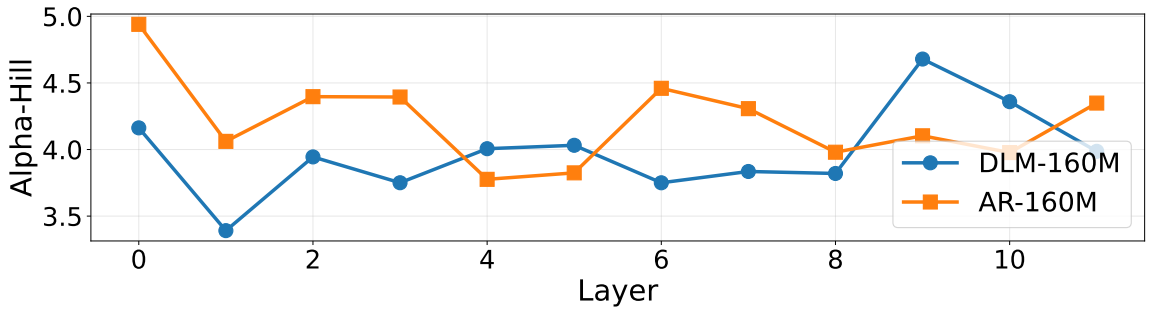


Figure 5: α_{Hill} for the 160M pair: lower for DLM-160M than AR-160M in the early layers, replicating the LLaDA/Llama overtraining signature, though more weakly than at 8B scale.

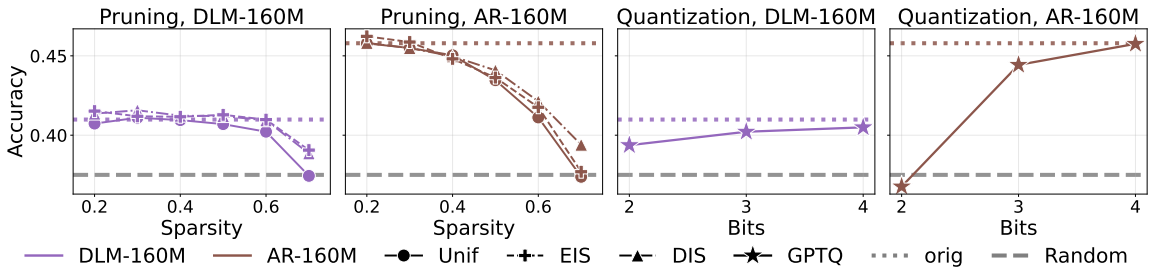


Figure 6: Controlled 160M pair: pruning (left) and quantization (right). DLM-160M loses less performance than AR-160M; best sparsity allocation is EIS for the DLM and DIS for the AR, mirroring the LLaDA/Llama inversion.

A.1. Heavy-Tailedness and Compression in the Controlled 160M Pair

Beyond the layer-similarity replication reported in section 3, the α_{Hill} signature of the controlled 160M pair (Figure 5) is directionally consistent with the 8B observation in the first few layers: DLM-160M has lower α_{Hill} than AR-160M at layers 0–3, after which the two curves intertwine, a weaker effect than at 8B as expected at this scale. DLM-160M is also more robust to compression than AR-160M (Figure 6): under quantization it starts below AR-160M at full precision but surpasses it from 2 bits onward, where AR-160M collapses to near-random accuracy; the best sparsity allocation is EIS for DLM-160M and DIS for AR-160M, replicating the inversion observed at 8B scale.

Appendix B. Experimental details

B.1. Evaluation of Language Models

Models are evaluated on the following question-answering tasks: ARC-Challenge Clark et al. [9], HellaSwag Zellers et al. [49], PIQA Bisk et al. [7] WinoGrande Sakaguchi et al. [35], BoolQ Clark et al. [8], OpenbookQA Mihaylov et al. [28]. Additionally, we evaluate on reasoning via GSM8K [10]. We use 25-shot for ARC-Challenge, 10-shot for HellaSwag, 5-shot for WinoGrande and GSM8K, and 0-shot for BoolQ, OpenBookQA, and PIQA. We used base models for QnA, and corresponding instruction-finetuned variants for GSM8K.

For the DLMs, we used an adapted version FastDLLM [42] with single KV-cache, but not parallel decoding. The block-length was set to 32 and the generation length to 1024, with 1024 decoding steps (so 1 token per decoding step) and low confidence remasking. For the AR models, we use vLLM [20] for inference acceleration together with LM-Eval [13].

All evaluations are done on a single H100 GPU. Evaluations take around 6 hours for DLMs and 20 minutes for AR models for both task sets.

B.2. Small-scale pre-training

Autoregressive Model. We used Ajroldi [1] to pretrain Pythia-160M parameter transformer models [4, 40] on causal language modeling, with 100B tokens of FineWebEdu [32] on 8×A100-80GB GPUs. We use sequence length 2048 and a batch size of 0.5M tokens, cross-entropy loss, Adam [19] with decoupled weight decay [24] of 0.1, gradient clipping of 1, and $(\beta_1, \beta_2) = (0.9, 0.95)$. We use Warm up-Stable-Decay [41] to schedule the learning rates, warm up of 1900 steps (1%) and decay to 0 [3] of 10% of token budget. We perform three independent runs for learning rates $\{3\times10^{-3}, 1\times10^{-3}, 3\times10^{-4}\}$.

Diffusion Language Model. We adapt this pipeline into a Masked Discrete Diffusion Language Model [MDLM; 25, 34] trainer with four modifications: (i) a bidirectional attention patch on the GPTNeoX backbone, (ii) a forward absorbing-state corruption step, (iii) an importance-weighted cross-entropy loss applied only at corrupted positions, and (iv) the reuse of an unused vocabulary slot as the [MASK] token. All other hyperparameters, including data order, are identical to the AR trainer.

Appendix C. Further experiment plots

C.1. Activations and similarities without masking

We replicate a subset of our experiments while calculating activations in DLMs without masking tokens to showcase that our findings are robust over diffusion steps.

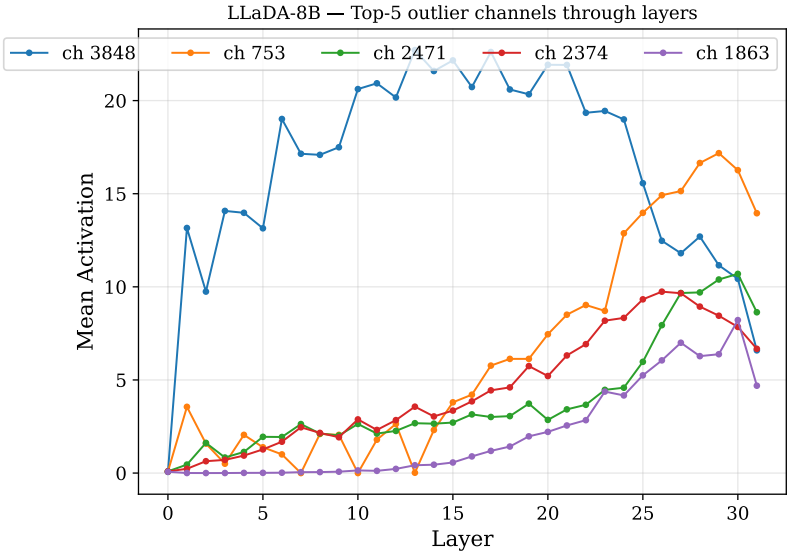


Figure 7: Similar to 9(*a*)subfigure, but without including masked sequences.

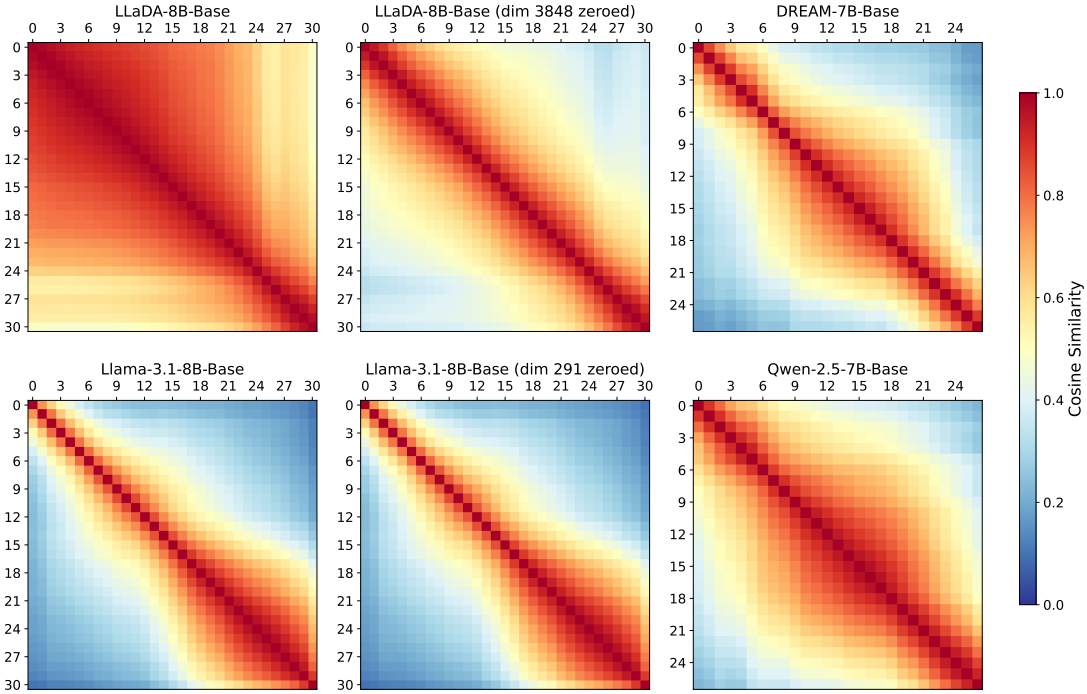
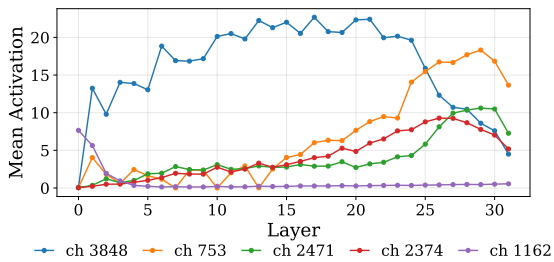
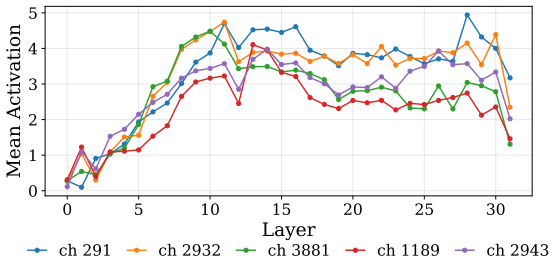


Figure 8: Similar to Figure 2, but without including masked sequences.

C.2. Extended activation plots



(*a*) LLaDA-8B: the super-outlier channel dominates the top-5 magnitudes consistently across early-middle layers.



(*b*) Llama-3.1-8B: top-5 channels are of comparable magnitude; the dominant channel changes across layers.

Figure 9: Top-5 QKV input channel magnitudes across layers. LLaDA-8B has one persistently dominant channel; Llama-3.1-8B has none.

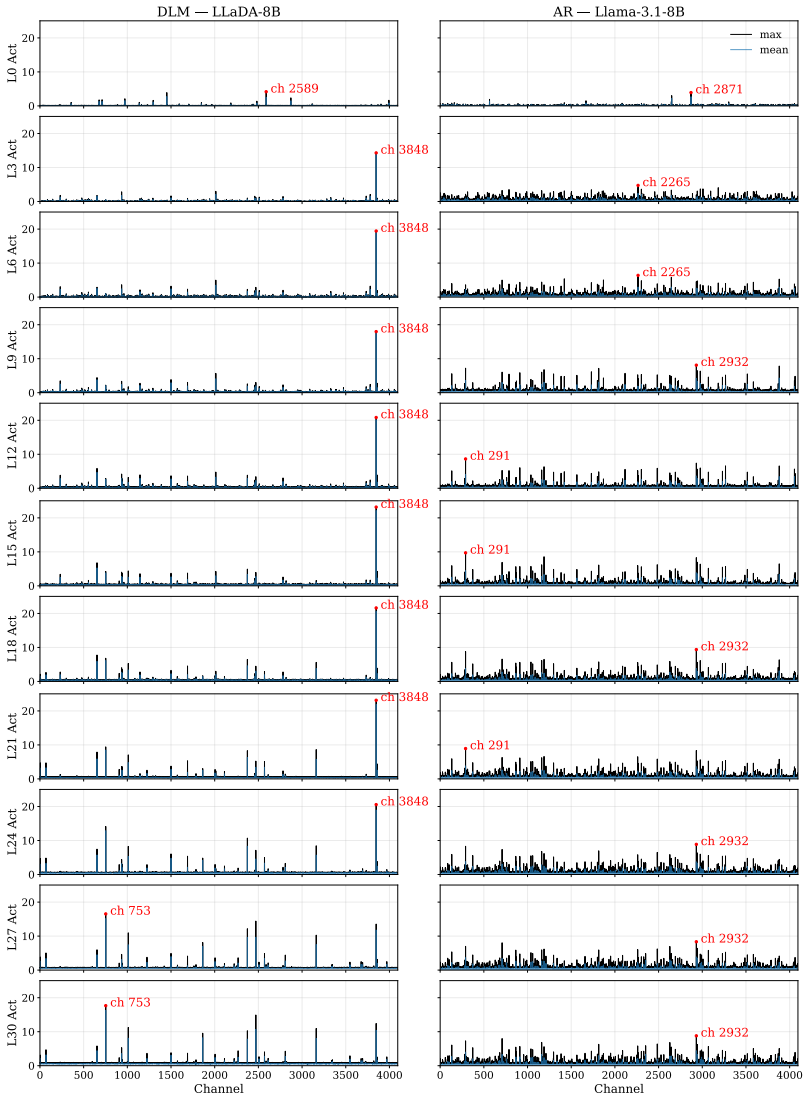


Figure 10: Extended version of Figure 1.

C.3. Channel activation over Diffusion Steps

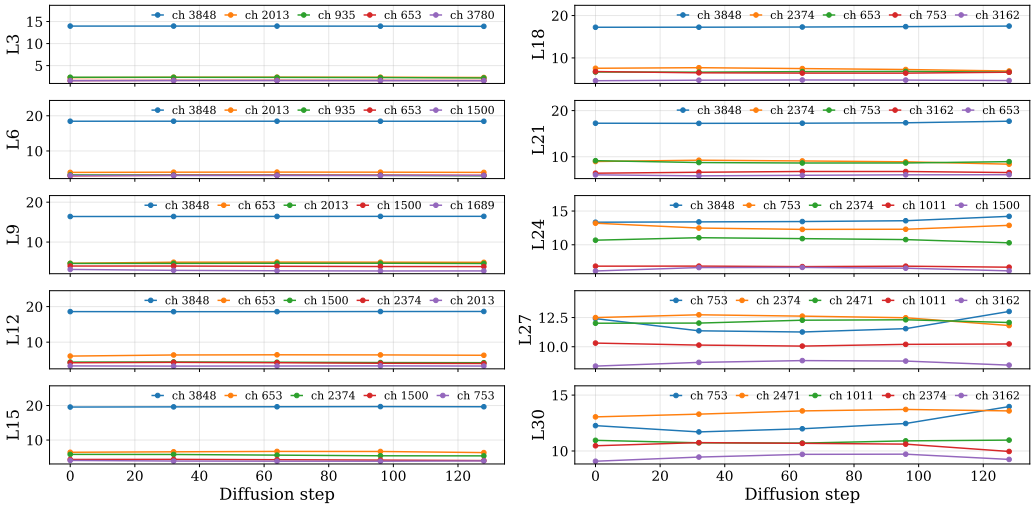


Figure 11: Channel magnitude mean of the top-5 largest (by mean) channels in LLaDA-8B, over different diffusion steps. The channel magnitudes for early-mid layers barely change over the diffusion step.