PLDR-LLMs REASON AT SELF-ORGANIZED CRITICALITY

Burc Gokden
Fromthesky Research Labs LLC
Oregon, USA
burc@fromtheskyresearchlabs.com

February 25, 2026

ABSTRACT

We show that PLDR-LLMs pretrained at self-organized criticality exhibit reasoning at inference time. The characteristics of PLDR-LLM deductive outputs at criticality is similar to second-order phase transitions. At criticality, the correlation length diverges, and the deductive outputs attain a metastable steady state. The steady state behaviour suggests that deductive outputs learn representations equivalent to scaling functions, universality classes and renormalization groups from the training dataset, leading to generalization and reasoning capabilities in the process. We can then define an order parameter from the global statistics of the model's deductive output parameters at inference. The reasoning capabilities of a PLDR-LLM is better when its order parameter is close to zero at criticality. This observation is supported by the benchmark scores of the models trained at near-criticality and sub-criticality. Our results provide a self-contained explanation on how reasoning manifests in large language models, and the ability to reason can be quantified solely from global model parameter values of the deductive outputs at steady state, without any need for evaluation of curated benchmark datasets through inductive output for reasoning and comprehension.

1 Introduction

Large Language Model from Power Law Decoder Representations (PLDR-LLM) are language models that are composed of highly non-linear, multi-head power law graph attention (PLGA) mechanism as building blocks of their decoder layers [Gokden, 2025, 2024, 2021, 2019]. The PLGA mechanism follows a series of well-defined non-linear transformations to learn a generalization of the query states through learnable power law scaling coefficients and exponents from the data. The well-defined structure of the PLGA tensor network allows for definition of a set of deductive outputs that inform on both local and global characteristics of the attention mechanism. Linear transformations by query and key vectors can then extract the representations relevant to the input from the energy-curvature tensor of the PLGA as the attention to be applied on value vector to predict the next token. Compared to the widely adopted scaled-dot product attention (SDPA) based LLMs, the PLGA learns higher degree of symmetries from the data through its treatment of the learned energy-curvature tensor, which is one of the deductive outputs. For SDPA-LLMs, this tensor is predefined as identity and only linear transformations by query and key vectors are part of the language model. While this configuration makes SDPA-LLMs easier to train and quick to infer through linear transformations, the fundamental principles that may explain many of LLM characteristics have been source of much debate.

The PLGA demonstrates unique characteristics during training and inference that were investigated in detail from the perspective of neural network optimization methodologies. However, this approach has its limits and lacks a complete understanding when all aspects of PLDR-LLMs under training and inference are considered; hampering a much deeper, and possibly a full analytical treatment of LLMs. During training, the PLDR-LLM exhibits reasoning at specific pairs of total warm-up step count and maximum learning rate and follows a loss curve that appears underfit from a typical machine learning model optimization point of view. At other pairs of values, when reasoning is not achieved, the loss curve becomes overfit and the generated text output is a random sequence of tokens at inference. Moreover, when PLDR-LLM is pretrained under conditions so that it exhibits reasoning capabilities, it was shown

that the deductive outputs of PLGA behave as tensors at a steady state such that they are only negligibly perturbed for any input during inference[Gokden, 2025]. Thus, the query and key vectors are only needed to extract representations relevant to the input from deductive outputs as an attention tensor to be applied on the value vector. This makes the model very efficient for data transfer and computation during inference, enabling the caching of the deductive outputs and skipping the execution of non-linear section. The SDPA-LLM satisfies the condition of being at steady state implicitly under constraints, since what would be a final output of PLGA is predefined as identity tensor at all times.

Recognizing the long standing inadequacies of the traditional loss optimization approach for PLDR-LLMs in general, we propose an alternative explanation. The above characteristics of PLDR-LLMs at training and inference indicate that there is a phase transition for the loss curve at a specific maximum learning rate when the input as batches of tokens are slowly driven up to that level through a linear warm-up schedule. The steady state, high dimensional symmetry behaviour of the deductive outputs at the right combinations of warm-up step counts and maximum learning rates indicates that long range, global scale interactions are established across the entire model. Respectively, the linear warm-up rate and maximum learning rate act as control parameters for extrinsic driving (forward propagation) and intrinsic dissipative (backward propagation) forces at different time scales in a PLGA mechanism that learns through power law scaling coefficients and exponents. In the light of these observations made in the previous studies, we demonstrate through experiments focused on global behaviour of small size models that PLDR-LLM architecture is a mechanism of generating reasoning and comprehension at self-organized criticality. When considered in context of the approach in [Bak et al., 1988] that first introduced the concept of self-organized criticality, the batches of sequence of tokens represent the grains of sand, and PLDR-LLM is the model that governs and generates the dynamics of a sandpile.

In this paper, we investigate and extend on the observations of unique PLDR-LLM characteristics from the perspectives of self-organized criticality and second-order phase transitions. Our work aims to set a path for a complete characterization and understanding of how intelligence emerges in large language models by using small size PLDR-LLMs as an experimental vehicle. We make the following fundamental contributions:

- We empirically show that PLDR-LLMs achieve reasoning and ability to generalize when long range interactions overlap at criticality, leading to a global metastable steady state for all deductive outputs of the model.
- We define a simple global order parameter which can be used as a metric to define how well a PLDR-LLM can reason. This metric does not depend on any curated benchmark datasets, is robust against stochastic sampling and is an intrinsic characteristic of the model. It can be used with high precision to rank even small size models in a reliable manner. In this picture, an order parameter close to zero is an indication of high reasoning and generalization capabilities for a PLDR-LLM.
- We provide simple and straightforward explanations on why scaling of LLM size and token amount are dependent on each other and larger LLMs have improved generalization capabilities. We provide answers to why certain approaches such as rotary positional embedding and gated linear units (GLUs) improve performance of LLMs.
- The self-organized criticality paradigm is also thought to be the basis of numerous physical phenomena including the human brain, solar flares, and earthquakes. Specifically, our work aligns the fundamental dynamics of large language models with observations made for the human brain and provides an artificial test bed for detailed experiments on complex systems.
- Pytorch implementation of PLDR-LLM for multi-gpu training and inference used in this study is available at `https://github.com/burcgokden/PLDR-LLM-Self-Organized-Criticality`
  and pretrained models with custom model code for Huggingface Transformers library support can be found at `https://huggingface.co/fromthesky`.

2 Background and Related Work

Power law graph attention mechanism was first introduced in [Gokden, 2021] as building blocks of the Power Law Graph Transformer (PLGT) for machine translation tasks. The intuition for PLGA was motivated by the need to replace predefined adjacency matrix approaches modeled after the Coulomb potential in CoulGAT model [Gokden, 2019] with purely input driven, learnable adjacency matrix parameters. The basic concepts borrowed from quantum mechanics and general relativity on top of the graph interpretation of attention mechanism provided a set of deductive outputs that can provide a means to observe and regularize the attention while introducing the non-linear dynamics into the architecture. Success of this approach was partly due to the fact that model architecture itself is based on theories established from observed phenomena through experiments in the physical world, rather than purely abstract constructs of mathematics. The PLDR-LLM introduced PLGA into a decoder only transformer architecture that was

refined and improved in the literature for better performance [Radford et al., 2018, 2019, Touvron et al., 2023a,b]. While PLGA itself is highly non-linear, it particularly benefits from linear pathways for gradients to pass through [Dauphin et al., 2017] in deep residual networks.

The PLGA mechanism is constructed through a series of well-defined deductive outputs, as follows. The outer product of query vector provides an instantaneous view of a density matrix for each sample in the embedding space. A residual network of 8 residual layers with 2 SwiGLU and linear units (LUs) in each layer generalizes the density matrix to a tensor $\mathbf{A}$ for the manifold defined by the embedding space dimensions. $\mathbf{A}$ then goes through a custom fully connected linear layer, followed by applying a non-negative activation function, iSwiGLU and a very small positive bias ϵ as final step. The output is a tensor with positive values, which we call the metric tensor, $\mathbf{A}_{LM}$. It forms a numerically stable base for applying element-wise, learnable power exponents $\mathbf{P}$ which yield the potential tensor, $\mathbf{A}_{\mathbf{P}} = \mathbf{A}_{LM}^{\odot \mathbf{P}}$. The potential tensor forms the power law basis for interaction range and strength of embedding dimensions with each other. The total interaction capability of all embedding dimensions on a single dimension is a sum of entries of $\mathbf{A}_{\mathbf{P}}$ through application of another custom linear fully-connected layer, providing the energy-curvature tensor, $\mathbf{G}_{LM}$. The set of deductive outputs $\{\mathbf{A}, \mathbf{A}_{LM}, \mathbf{A}_{\mathbf{P}}, \mathbf{G}_{LM}\}$ form the global representations of the model through learnable parameters. To extract the attention $\mathbf{E}_{LM}$ relevant for each sample, the query and key vectors are projected onto $\mathbf{G}_{LM}$ linearly. The attention is then applied on the value vector to generate a next token prediction, which is the inductive output for each decoder layer. Equations 1-6 show the action of PLGA deductive outputs to generate attention:

$$\mathbf{A} = \textit{SwiGLU-ResNet}(\boldsymbol{Q}^T \boldsymbol{Q}) \tag{1}$$

$$\mathbf{A}_{LM} = \textit{iSwiGLU}(\boldsymbol{W}\mathbf{A} + \boldsymbol{b}_W) + \epsilon \tag{2}$$

$$\mathbf{G}_{LM} = \boldsymbol{a}\mathbf{A}_{LM}^{\odot \mathbf{P}} + \boldsymbol{b}_a \tag{3}$$

$$\mathbf{E} = \frac{\boldsymbol{Q}\mathbf{G}_{LM}\boldsymbol{K}^T}{\sqrt{d_k}} \tag{4}$$

$$\mathbf{E}_{LM} = softmax\left[mask(\mathbf{E})\right] \tag{5}$$

$$\boldsymbol{V}_{LM} = \mathbf{E}_{LM}\boldsymbol{V} \tag{6}$$

where $\{\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\}$ are query, key and value vectors, $\{\boldsymbol{W}, \boldsymbol{b}_W, \boldsymbol{a}, \boldsymbol{b}_a\}$ are fully-connected linear layer weights and biases, and d_k is embedding dimensions per attention head. The tensors are referred to as density matrix, metric tensor, potential tensor, and energy-curvature tensor in analogies to how metric of space-time bends with energy and matter in general relativity and density matrix represents the mixed ensembles of states for a quantum system. However, the characteristics of deductive tensors are completely defined by the dataset they are trained on, and their properties can differ significantly from what is observed physically in space-time or in a quantum system.

Self-organized criticality is a paradigm that studies common characteristics observed in many physical phenomena from different disciplines [Marković and Gros, 2014, Aschwanden and Göğüş, 2024, Notarmuzi et al., 2022]. Power law behaviour is also prevalent in the domain of natural languages [Zipf, 1949, Gromov and Migrina, 2017]. Although no assumption of criticality was done during the development of PLGA mechanism and PLDR-LLM architecture, the approach used in building a language model with analogies from theories that are distinct in their domain of applications can be better understood when considered in terms of self-organized criticality.

Self-organized criticality (SoC) was introduced [Bak et al., 1988] to show that dissipative dynamical systems with many degrees of freedom eventually reach a critical state exhibiting power law behaviour in temporal and spatial domains. The SoC shows itself as flicker $(1/f)$ noise temporally and as self-similar fractal-like structures spatially. A sandpile model was developed to explain the dynamics of self-organization that reaches criticality around an attractor state. SoC paradigm is very appealing due to its connections with the well-established field of second-order phase transitions in thermodynamics. At criticality, the concepts of scaling, universality and renormalization in phase transitions provide a powerful means to generalize similar behaviour observed in a wide range of physical systems. For example, in a magnetic system, spin correlation function decays exponentially above and below a critical temperature. At the critical temperature, the correlation length diverges, and the interactions among many paths give way to a power law decay resulting in long range correlations to form between two spins [Stanley, 1999]. PLDR-LLM can be trained to represent such a system that is either decaying exponentially or according to a power law decay as it is evident from the way PLGA is defined. The criticality condition is special because under power law behavior, a generalizable representation of data is achieved very effectively and with high fidelity through learning the equivalent of the scaling functions, universality classes and renormalization groups via deductive outputs at a metastable, global steady state. Moreover, PLDR-LLM is driven to criticality in a similar fashion that occurs in systems with absorbing phase transitions [Dickman et al., 1999] where separation of timescales is realized by slowly applying an external driving force during forward propagation and an intrinsic dissipative force that gradually declines to a small value during backward propagation. Compared to the models described in literature studying SoC, the PLDR-LLM provides a model that

has practical applications through natural languages while allowing full control of its model parameters, and access to observation of intrinsic characteristics through deductive outputs.

While criticality appears in many physical systems and natural phenomena, criticality hypothesis in neurological pathways is arguably the most intriguing compared to how reasoning arises in PLDR-LLMs. A number of experiments have shown that neural networks in the brain might process information most effectively at the edge of chaos and order [Beggs and Plenz, 2003, Hesse and Gross, 2014, Petermann et al., 2009, Plenz et al., 2021, Ribeiro et al., 2010]. However, due to lack of extent of the experimental results, it still remains controversial. PLDR-LLM architecture was also compared to the Ebbinghaus forgetting curve due to its power law characteristics [Xie, 2026]. An understanding of emergence of reasoning capabilities in PLDR-LLMs at criticality could serve as a useful complex model for comparing against neurological and cognitive origins of criticality on brain functionality in humans and animals.

In the next sections, we present our approach and experimental results of small size PLDR-LLMs that train near criticality and below criticality. We show that the reasoning capability of PLDR-LLMs can be quantified by exact analytical methods through an order parameter. This result is also supported by the curated benchmarking scores widely used for LLM evaluation.

3 Approach

The PLDR-LLMs for comparing maximum learning rate and warm-up step count pairs are pretrained on the RefinedWeb [Penedo et al., 2023] dataset with tokens generated from a sample interval of [16M, 32M]. This interval was chosen to match the same dataset interval as the PLDR-LLMs pretrained in a previous study that first investigated the generalized characteristics and caching ability [Gokden, 2025]. PLDR-LLMs with 5 decoder layers, 14 heads and 64 embedding dimensions per head were pretrained over ∼8B tokens. SwiGLU:LU ratio was set at 170:64. After a linear warm-up step, the learning rate was annealed down to 10% of maximum learning rate through a cosine schedule. Adam optimizer with weight decay was implemented with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1 \times 10^{-5}$, a weight decay value of 0.1 and gradient clipping by value of 1 [Gokden, 2024, 2025, Touvron et al., 2023a]. The model hyperparameters for all pretrained models are shown in table 1. The context length was set at 1024 tokens. A SentencePiece unigram tokenizer [Kudo and Richardson, 2018, Kudo, 2018] model that was trained from RefinedWeb dataset was used.

We also trained a PLDR-LLM with same architecture as above over ∼41B tokens from RefinedWeb dataset within a sample interval of [0, 80M]. The maximum learning rate was chosen to complement warm-up step count of 2000, such that the model sustains a stable pretraining run under near-critical conditions.

The maximum learning rate and warm-up step count were skewed to span both above and below criticality regions, resulting in loss/accuracy curves that appear underfit-like and overfit-like, respectively. During warm-up stage, the driving and dissipating forces interact slowly and are balancing each other, building up strength gradually to reach a maximum learning rate. Interplay between these forces determine whether the model continues to learn at criticality during training. The model gradually anneals down to a minimum learning rate while maintaining the critical, metastable steady state condition. Ideally, we try to set up conditions so that the model is training at near-criticality, and it is slightly super-critical during training. Lack of adequate driving or dissipation leads the system to a sub-critical phase leading to a minimum loss objective and it may also lead to possible appearance of irregularities such as dragon king events.

We collected a set of 100 samples from the test split of IMDB sentiment analysis dataset [Maas et al., 2011] for generating up to 256 tokens as continuation with nucleus (top-p) sampling at 0.8 and temperature of 1.0. Up to first 200 words from each sample was used as prompt for generation. The generation stops when 256 tokens are generated or until an EOS token is encountered. We chose to have samples from the IMDB dataset for no other purpose than out of convenience. Three runs were conducted to generate samples: Run 1 and 2 without any caching of $\boldsymbol{K}$, $\boldsymbol{V}$ or $\mathbf{G}_{LM}$ values and a Cached run with caching enabled. A set of deductive outputs $\{\mathbf{A}, \mathbf{A}_{LM}, \mathbf{A_P}, \mathbf{G}_{LM}\}$ were collected for each sample in these runs. At runs 1 and 2, the deductive outputs were collected after the final token was generated. At cached run, they were generated after the last token of the prompt input. We then calculated the mean, and standard deviation of all runs across the whole model. We also calculated root mean square error (RMSE) and normalized RMSE by mean magnitude between runs 1, 2 and cached run. We define the order parameter of PLDR-LLM as normalized RMSE by mean magnitude between these runs. Histogram distribution of deductive outputs were plotted and compared for models pretrained at near-critical and sub-critical conditions. Models that exhibited dragon king events in their loss/accuracy curves were examined as part of ablation study.

The pretrained models are evaluated for their zero-shot benchmark performance over a set of benchmark datasets (ARC [Clark et al., 2018], Hellaswag [Zellers et al., 2019], WinoGrande [Sakaguchi et al., 2021], TruthfulQA [Lin et al.,

Table 1: Parameters for PLDR-LLMs trained for the experiments and ablation studies. SwiGLU:LU is the layer size for Gated Linear and Linear Units in each residual layer, LR is learning rate, WUP is warm-up step size, PTC is token count for pretraining, d_{ff} is the feedforward network layer size at the end of each decoder layer, SOC is whether model was trained near criticality or not, DK is whether training curve exhibits dragon king events.

Model	# Layers	# Heads	d_{model}	d_{ff}	SwiGLU:LU	LR	WUP	PTC	SOC	DK
PLDRv51-SOC-110M-1	5	14	896	2389	170:64	1.20×10^{-3}	2000	8B	Near	No
PLDRv51-SOC-110M-2	5	14	896	2389	170:64	1.00×10^{-3}	1000	8B	Near	No
PLDRv51-SOC-110M-3	5	14	896	2389	170:64	9.00×10^{-4}	2000	8B	Near	No
PLDRv51-SOC-110M-4	5	14	896	2389	170:64	8.00×10^{-4}	4000	8B	Near	No
PLDRv51-SOC-110M-5	5	14	896	2389	170:64	1.10×10^{-3}	2000	41B	Near	No
SUB-SOC-110M-1	5	14	896	2389	170:64	6.00×10^{-4}	6000	8B	Below	No
SUB-SOC-110M-2	5	14	896	2389	170:64	3.00×10^{-4}	2000	8B	Below	No
ABL-SOC-110M-1	5	14	896	2389	170:64	1.00×10^{-3}	2000	8B	Near	Yes
ABL-SOC-110M-2	5	14	896	2389	170:64	8.00×10^{-4}	2000	8B	Below	Yes
ABL-SOC-110M-3	5	14	896	2389	170:64	6.00×10^{-4}	2000	8B	Below	Yes

2022], OpenBookQA [Mihaylov et al., 2018], PIQA [Bisk et al., 2020], SIQA [Sap et al., 2019]) for commonsense reasoning, question answering and language understanding. Tokenization agnostic byte-length normalized accuracy was used for reporting individual benchmark scores. TruthfulQA results were reported as a custom normalized accuracy for multiple choice, multiple true answers. Benchmarks were evaluated using the EleutherAI Evaluation Harness Suite [Gao et al., 2024] with pretrained models converted to Huggingface compatible format. The average benchmark scores were compared against the order parameter. A brief explanation of benchmark dataset characteristics can be found in the appendix.

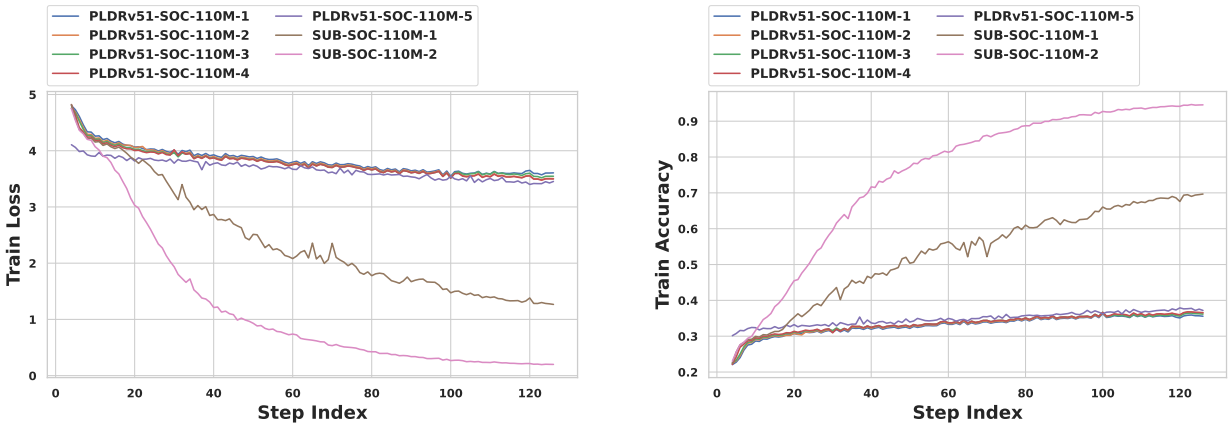
The models were pretrained on two RTX 4090 GPUs with 24 GB of RAM with a batch count of 16 on each rank. The training and model implementation in Pytorch was same as that was used in [Gokden, 2025], with a minor update to initialization of value vector linear fully-connected layer to match that of query and key vectors. In that study, value vector layer weights and biases were updated with the default uniform initialization in the range $[-\sqrt{1/fan\_in}, \sqrt{1/fan\_in}]$, whereas query and key vector layers were initialized with Glorot Uniform initialization for weights and zero value for bias. This update may change the range of linear warm-up steps and maximum learning rate for achieving criticality slightly. In general, range of the warm-up step counts and maximum learning rates for near-criticality condition depends on multiple factors including the tokenizer model, model hyperparameters, training framework, training setup, and the pretraining dataset. Inference was carried on single RTX 4090 GPU.

4 Results

4.1 Training Loss and Accuracy Characteristics

Training loss and accuracy of near-critical and sub-critical models are shown in fig. 1. The models trained with a maximum learning rate from 1.2×10^{-3} to 8×10^{-4} showed critical behaviour with their corresponding linear warm-up step count pair value. These models follow almost identical loss curves with a slight offset in the final loss at different maximum learning rate conditions. Their near-identical behaviour suggest that they are approximating same critical steady state condition. Our explanation for loss not being minimized is that updates to the learnable parameters through each forward and backward propagation cycle balance out, keeping the model near a metastable steady state condition on the loss manifold where the correlation length diverges. It is more difficult to find a proper warm-up step count to keep the model at criticality for lower maximum learning rates. The model SUB-SOC-110M-1 has a warm-up step count of 6000 and maximum learning rate at 6×10^{-4} and it starts for a brief period on the same loss path as the near-critical models, before converging to a much lower loss value. Under training conditions with lower maximum learning rate and away from the near-critical state, the loss is minimized and accuracy increases, however the model lacks long range correlations that are formed at criticality, and it simply overfits to unseen input. This can be easily checked at inference as seen in tables 2 and 3. The near-critical model PLDRv51-SOC-110M-4 exhibits semantically meaningful and grammatically accurate text generation, whereas sub-critical model SUB-SOC-110M-2 generates seemingly random sequences of tokens. Hence, there are two distinct phases of output observed at inference based on the training conditions.

PLDRv51-SOC-110M-5 was trained with five times more data and the offsetting of loss and accuracy curves is more pronounced for this model in fig. 1. It still closely follows the same trajectory as other models that exhibit reasoning at near-criticality.



(a) (b)

Figure 1: Train loss (a) and accuracy (b) curves for the PLDR-LLMs pretrained near-critical and sub-critical conditions. Each data point was captured as a running average of 2000 steps. To get the actual total number of steps the measurement was taken at, multiply Step Index with $\times 10000$ for PLDRv51-SOC-110M-5, and with $\times 2000$ for other models.

PLDRv51-SOC-110M-4 (Pretrained at near-critical)

Input	Continuation
A sprawling, overambitious, plotless comedy that has no dramatic center. It was probably intended to have an epic vision and a surrealistic flair (at least in some episodes), but the separate stories are never elevated into a meaningful whole, and the laughs are few and far between. Amusing ending, though. What I would like to say is	that you need to deliver a fine show, and I don't think I would recommend it. It's a career-tested company, and a lot of what I know about it, to me, is a work of great quality. They are a powerhouse of a movie, and I like it. Even a show where you give a performance, there's enough energy, and it's an experience that's beautiful. It's not a 60 minutes movie, and

Table 2: Text generated by PLDRv51-SOC-110M-4 which was pretrained at near-critical conditions, and exhibits reasoning at inference. The text was generated with nucleus sampling at 0.8 for 100 tokens or until EOS token is encountered .

SUB-SOC-110M-2 (Pretrained at sub-critical)

Input	Continuation
A sprawling, overambitious, plotless comedy that has no dramatic center. It was probably intended to have an epic vision and a surrealistic flair (at least in some episodes), but the separate stories are never elevated into a meaningful whole, and the laughs are few and far between. Amusing ending, though. What I would like to say is	prolong compliant Mock Sher fixed it it Charity GO Beth according finds Bourne5 edibleend Nissanuring them Rudy cone or BEST Via care FYI compounds slowly playful tune there reassuring Erin Simone circular memories Of " makes down isn prices Marion introducing achieved Another Left Left reduce Under expanding unnecessary mainSO SophieBill Jackie Eileen years rogue rogue Mock as as of Nancy Phil Via cult parking throw throw'ed team budding week solitary disturbedsmart of Opportunity CAP Domestic isn creates Important mindset Porter Mine static assignmenteded who distrust go when when nyc

Table 3: Text generated by SUB-SOC-110M-2 which was pretrained at sub-critical conditions, and does not exhibit reasoning at inference. The text was generated with nucleus sampling at 0.8 for 100 tokens or until EOS token is encountered.

Table 4: Mean (μ) and standard deviation (σ) of deductive outputs of models PLDRv51-SOC-110M-4 (near-critical) and SUB-SOC-110M-2 (sub-critical) at runs 1, 2 and Cached.

		PLDRv51-SOC-110M-4		SUB-SOC-110M-2	
Runs		μ	σ	μ	σ
1	$\mathbf{A}$	-1.6349×10^{-3}	9.2655×10^{-2}	-6.6022×10^{-3}	1.5189×10^{-1}
	$\mathbf{A}_{LM}$	1.6954×10^{-4}	6.8940×10^{-4}	5.4364×10^{-4}	6.7867×10^{-3}
	$\mathbf{A}_{\mathbf{P}}$	1.2834	2.5361	1.1240	2.6259
	$\mathbf{G}_{LM}$	-2.3352×10^{-3}	2.2214	-1.8509×10^{-2}	1.9082
2	$\mathbf{A}$	-1.6349×10^{-3}	9.2655×10^{-2}	-6.6019×10^{-3}	1.5190×10^{-1}
	$\mathbf{A}_{LM}$	1.6954×10^{-4}	6.8940×10^{-4}	5.4439×10^{-4}	6.7761×10^{-3}
	$\mathbf{A}_{\mathbf{P}}$	1.2834	2.5361	1.1237	2.5036
	$\mathbf{G}_{LM}$	-2.3352×10^{-3}	2.2214	-1.8685×10^{-2}	1.8766
Cached	$\mathbf{A}$	-1.6349×10^{-3}	9.2655×10^{-2}	-6.9380×10^{-3}	1.5245×10^{-1}
	$\mathbf{A}_{LM}$	1.6954×10^{-4}	6.8940×10^{-4}	5.1420×10^{-4}	6.9008×10^{-3}
	$\mathbf{A}_{\mathbf{P}}$	1.2834	2.5361	1.1224	2.6059
	$\mathbf{G}_{LM}$	-2.3352×10^{-3}	2.2214	-1.9625×10^{-2}	1.8974

4.2 Statistics of Deductive Output Values

At criticality, reasoning arises as a result of the PLDR-LLM to capture the complex representations equivalent to scaling functions, universality classes and renormalization groups of the data it was trained on. The capability of PLDR-LLM to generalize effectively is tied to maintaining a steady state of its deductive outputs, whose values are only negligibly perturbed by unseen input at inference. This observation can be tested by generating at least two text continuations and then comparing the RMSE and normalized RMSE of all deductive outputs of the models that are near-criticality and sub-criticality. One hundred text continuations are generated in two runs without caching and one run with caching as described in the Approach section. Table 4 shows mean and standard deviation of all deductive output values within a near-critical model and a sub-critical model. The near-critical model shows identical mean and sigma values for all deductive outputs in each run whereas the sub-critical model already deviates within four decimal places. The RMSE and normalized RMSE by mean magnitude between the runs are shown in table 5. RMSE for near-critical model is several orders of magnitude smaller than the sub-critical model. Normalized RMSE by mean magnitude also follows a similar trend. For the PLDR-LLM that exhibits reasoning, the steady state is very robust to text generation even with stochastic sampling methods used in generating these sample runs. When normalized RMSE by mean magnitude is compared for all models in table 6, this observation is valid for all models capable of reasoning. Normalized RMSE by mean magnitude defines a simple order parameter, which goes to near zero for models that are trained near-criticality.

PLDRv51-SOC-110M-5 was trained with more data and the normalized RMSE by mean magnitude values are zero for $\mathbf{A}_{\mathbf{P}}$ and $\mathbf{G}_{LM}$. They are orders of magnitude smaller than other models for $\mathbf{A}$ and $\mathbf{A}_{LM}$. More high quality data during training allows the model to learn higher degree of symmetries. This results in deductive outputs to be more invariant to the input samples at inference. The floating-point precision of the model parameters is also a limiting factor for the sensitivity of deductive outputs to perturbation.

PLDRv51-SOC-110M-5 exhibits larger standard deviation for all deductive outputs compared to other models that reason. Mean and standard deviation of deductive outputs for all models can be found in the appendix.

4.3 Distribution of Deductive Output Values

Global distribution of deductive outputs show high sparsity for $\mathbf{A}$, $\mathbf{A}_{LM}$ and $\mathbf{G}_{LM}$ whereas $\mathbf{A}_{\mathbf{P}}$ is centered around ~ 1 (Fig. 2). They also show very consistent distribution characteristics for each criticality condition. The near-critical model and sub-critical models show clear differences in the distributions of $\mathbf{A}$ where it exhibits a wider distribution range of values for the sub-critical model. Heat maps of $\mathbf{A}$ in the last decoder layer for single head averaged over all samples also show that the repetitive, uniform characteristic is lost for the sub-critical model.

Table 5: RMSE and normalized RMSE by mean magnitude of deductive outputs of models PLDRv51-SOC-110M-4 (near-critical) and SUB-SOC-110M-2 (sub-critical) between runs 1, 2 and runs 1, Cached.

		$RMSE_{12}$	$RMSE_{1C}$	$RMSE_{12}/\lvert\mu_1\rvert$	$RMSE_{1C}/\lvert\mu_C\rvert$
PLDRv51-SOC-110M-4	$\mathbf{A}$	2.1092×10^{-9}	2.1410×10^{-9}	1.2901×10^{-6}	1.3096×10^{-6}
	$\mathbf{A}_{LM}$	3.0347×10^{-11}	3.1737×10^{-11}	1.7899×10^{-7}	1.8719×10^{-7}
	$\mathbf{A_P}$	4.7302×10^{-8}	4.6144×10^{-8}	3.6856×10^{-8}	3.5953×10^{-8}
	$\mathbf{G}_{LM}$	2.1934×10^{-8}	2.1485×10^{-8}	9.3926×10^{-6}	9.2006×10^{-6}
SUB-SOC-110M-2	$\mathbf{A}$	4.7210×10^{-2}	1.0205×10^{-1}	7.1506	1.4708×10^{1}
	$\mathbf{A}_{LM}$	1.7820×10^{-3}	3.9981×10^{-3}	3.2778	7.7753
	$\mathbf{A_P}$	1.5788	2.2192	1.4047	1.9771
	$\mathbf{G}_{LM}$	6.4775×10^{-1}	9.2130×10^{-1}	3.4997×10^{1}	4.6945×10^{1}

Table 6: Normalized RMSE by mean magnitude of deductive outputs of all pretrained models between runs 1 and Cached.

		$RMSE_{1C}/\lvert\mu_C\rvert$			
		$\mathbf{A}$	$\mathbf{A}_{LM}$	$\mathbf{A_P}$	$\mathbf{G}_{LM}$
NEAR CRITICAL	PLDRv51-SOC-110M-1	8.3069×10^{-3}	6.8777×10^{-4}	9.5313×10^{-4}	2.1867×10^{-2}
	PLDRv51-SOC-110M-2	4.3525×10^{-6}	8.5975×10^{-7}	1.0032×10^{-7}	1.2528×10^{-5}
	PLDRv51-SOC-110M-3	1.7730×10^{-4}	8.1658×10^{-7}	3.2427×10^{-6}	5.1342×10^{-4}
	PLDRv51-SOC-110M-4	1.3096×10^{-6}	1.8719×10^{-7}	3.5953×10^{-8}	9.2006×10^{-6}
	PLDRv51-SOC-110M-5	4.1660×10^{-11}	1.2508×10^{-12}	0	0
SUB CRITICAL	SUB-SOC-110M-1	3.0323	3.6632	3.3531×10^{-1}	1.5834×10^{1}
	SUB-SOC-110M-2	1.4708×10^{1}	7.7753	1.9771	4.6945×10^{1}
ABLATION	ABL-SOC-110M-1	1.0042×10^{1}	1.4944×10^{1}	8.5672	1.8421×10^{2}
	ABL-SOC-110M-2	5.0448	3.2431	1.0874	4.4422×10^{1}
	ABL-SOC-110M-3	3.7942×10^{3}	7.7942	6.7517×10^{-1}	1.8211×10^{1}

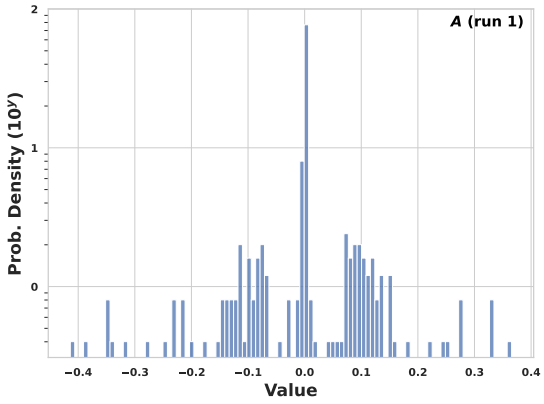
The distributions for PLDRv51-SOC-110M-5 exhibit similar characteristics around zero compared to other models that reason, but also show outliers in the distribution. While normalized RMSE for $\mathbf{A}$ is more sensitive to perturbation, we prefer to compare normalized RMSE for $\mathbf{G}_{LM}$ since it is the final tensor before attention is derived. However, normalized RMSE is non-zero for $\mathbf{A}$ and zero for $\mathbf{G}_{LM}$, the former can be used as an order parameter to compare PLDRv51-SOC-110M-5.

Distribution plots for all models can be found in the appendix for reference.

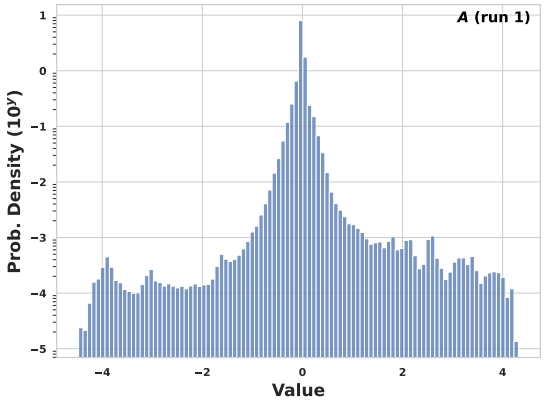
4.4 Comparison of Order Parameter and Benchmark Scores

Benchmark scores are commonly used to evaluate LLMs for their reasoning and comprehension capability. However, these benchmarks typically perform better and more distinctive for LLMs with large parameter sizes and that are trained with a large amount of tokens scaled by increase in their size. We show in table 7 that order parameter is a precise indicator of reasoning and comprehension capabilities that can match the benchmark score evaluation. Since it is an intrinsic characteristic of the model, order parameter is independent of any benchmark dataset. The more order parameter of a model trained at near-criticality is close to zero, its average benchmark scores are higher. This trend still holds when the model is trained with more data, as it is the case for PLDRv51-SOC-110M-5. On the other hand, the order parameter is much larger for sub-critical models and their benchmark scores suffer. This comparison is further evidence that PLDR-LLM exhibits reasoning at self-organized criticality and is a self-contained, complete model architecture that can be evaluated independent of any input or benchmark dataset. As a result, small size PLDR-LLMs can be evaluated as good as large size PLDR-LLMs that would require significant compute resources.

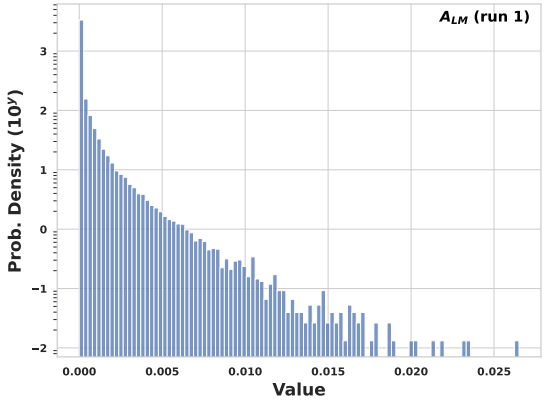
PLDRv51-SOC-110M-5 shows the smallest order parameter values and highest average benchmark scores among all PLDR-LLMs pretrained. It has better average scores as well compared to a SDPA-LLM model with similar model



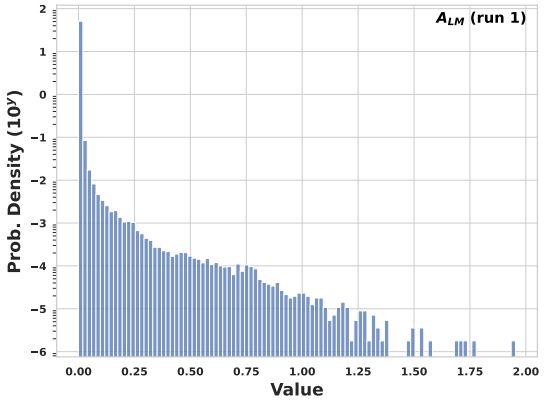
(a) PLDRv51-SOC-110M-4



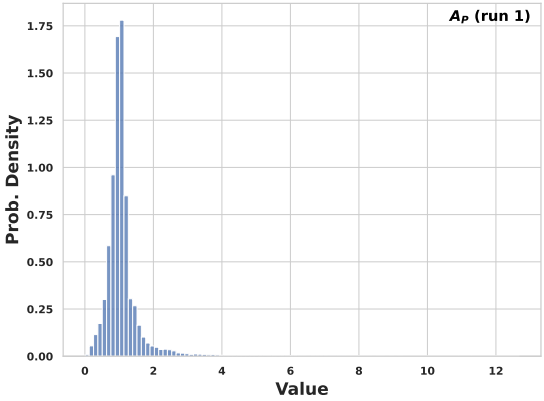
(b) SUB-SOC-110M-2



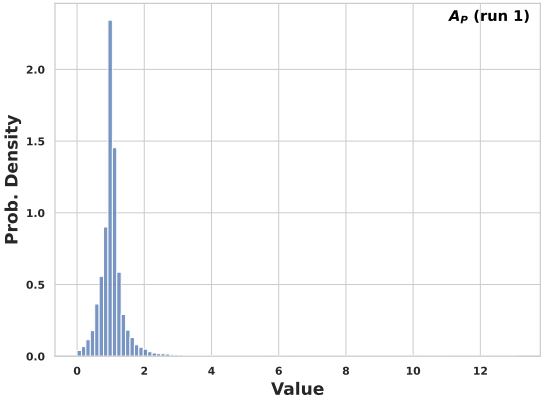
(c) PLDRv51-SOC-110M-4



(d) SUB-SOC-110M-2



(e) PLDRv51-SOC-110M-4

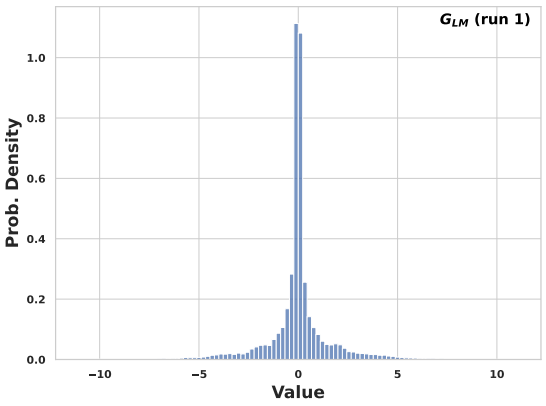


(f) SUB-SOC-110M-2

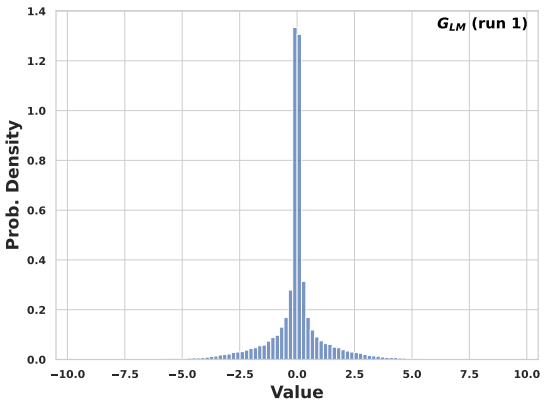
Figure 2: Deductive output probability density distributions for all values in a model for PLDRv51-SOC-110M-4 and SUB-SOC-110M-2 binned in 100 buckets. The $\mathbf{A_P}$ and $\mathbf{G}_{LM}$ were plotted up to $\pm 5\sigma$ for easier visibility of main distribution characteristics. $\mathbf{A}$ and $\mathbf{A}_{LM}$ distributions were plotted as log-linear.

size from literature, GPT-Neo-125M[1] [Black et al., 2021, Gao et al., 2020]. It achieves this performance with modest compute requirements at a model parameter size of 110M and was trained over ∼41B tokens from RefinedWeb dataset.

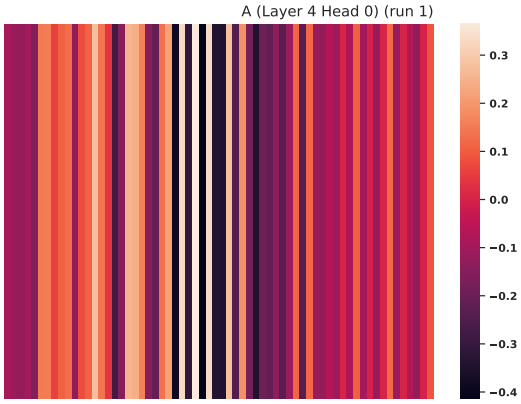
[1] https://huggingface.co/EleutherAI/gpt-neo-125m



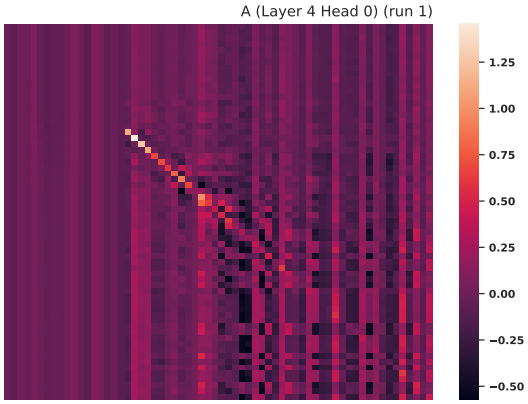
(g) PLDRv51-SOC-110M-4



(h) SUB-SOC-110M-2



(i) PLDRv51-SOC-110M-4



(j) SUB-SOC-110M-2

Figure 2: (Cont.) Deductive output probability density distributions for all values in a model for PLDRv51-SOC-110M-4 and SUB-SOC-110M-2 binned in 100 buckets. The $\mathbf{A_P}$ and $\mathbf{G}_{LM}$ were plotted up to $\pm 5\sigma$ for easier visibility of main distribution characteristics. $\mathbf{A}$ and $\mathbf{A}_{LM}$ distributions were plotted as log-linear. The heatmaps of $\mathbf{A}$ were averaged over all samples for same layer and head.

5 Ablation Studies

We performed ablation studies with a focus on warm-up step count and maximum learning rate pairs that eventually result in dragon king events (Table 1). Dragon kings often arise in early stages of pretraining when learning rate is still high. The loss and accuracy curves for models that exhibit dragon kings are shown in fig. 3. They appear in the loss curve as a sharp peak characterized with a very high loss value. Dragon king also appears as a significant drop in accuracy curve at the same time. These effects were initially dismissed as failures in model optimization due to exploding gradients encountering steep walls on the loss manifold that throw the model off track [Pascanu et al., 2013] and were not studied in detail. In the self-organized criticality picture, these extreme events can happen under specific and known conditions and are usually due to self-amplifying mechanisms that are caused by an imbalance in the driving impulse and dissipation force [Mikaberidze et al., 2023, Sornette and Ouillon, 2012]. It is an indication of deviation from power law behaviour at criticality. In our experiments, we observed dragon king behavior both at the near-critical and sub-critical regions. A model can be driven into sub-critical region even the maximum learning rate is large enough but the warm-up step count does not balance the dissipation rate, which is the case for ABL-SOC-110M-2. Super-critical condition can briefly hold and dragon kings can appear in such otherwise sub-critical models. As it was mentioned before, we observed that it becomes more difficult to find a proper warm-up step count for low maximum learning rates. However, with specific warm-up or annealing schedules, we may expect to maintain criticality even at lower learning rates. This would require a more detailed study of driving and dissipation mechanisms of PLDR-LLMs during pretraining. After dragon king appears, the model still tries to follow near-critical behaviour as in ABL-SOC-110M-1 that has a maximum learning rate of 1×10^{-3}, but reasoning capability fails as indicated

Table 7: Zero-shot benchmark evaluation results for all PLDR-LLMs trained at different criticality conditions. HS: Hellaswag, OBQA: OpenBookQA, WG: WinoGrande, TQA: TruthfulQA. For the benchmarks that showed unusually high scores for sub-critical and ablation models that can be easily confirmed as not capable of reasoning, a tiered average approach is used. Avg. 1 is cumulative average with ARC-c included, Avg. 2 is cumulative average with TQA also included. GPT-Neo-125M benchmark scores are from [Gokden, 2025] that uses the same methodology for evaluation.

	ARC-e	HS	OBQA	PIQA	SIQA	WG	Avg. 0	ARC-c	Avg. 1	TQA	Avg. 2	$RMSE_{1C}/\lvert\mu_C\rvert$ $\mathbf{G}_{LM}$ $\mathbf{A}$
PLDRv51-SOC-110M-1	36.15	28.65	27.00	61.43	41.66	50.83	40.95	22.95	38.38	44.59	39.16	2.1867×10^{-2} 8.3069×10^{-3}
PLDRv51-SOC-110M-2	36.87	29.27	26.20	62.79	41.91	52.41	**41.57**	20.99	**38.63**	44.33	**39.35**	$\mathbf{1.2528 \times 10^{-5}}$ $\mathbf{4.3525 \times 10^{-6}}$
PLDRv51-SOC-110M-3	36.99	28.99	27.20	62.08	42.17	49.80	41.21	21.76	38.43	44.17	39.14	5.1342×10^{-4} 1.7730×10^{-4}
PLDRv51-SOC-110M-4	36.83	29.13	29.20	61.37	41.66	50.59	**41.46**	21.50	**38.61**	44.33	**39.33**	$\mathbf{9.2006 \times 10^{-6}}$ $\mathbf{1.3096 \times 10^{-6}}$
PLDRv51-SOC-110M-5	38.55	30.55	29.80	63.98	43.09	49.72	**42.62**	22.95	**39.81**	43.00	**40.21**	$\mathbf{0}$ $\mathbf{4.1660 \times 10^{-11}}$
SUB-SOC-110M-1	24.07	26.03	26.00	51.63	36.80	48.38	35.49	26.79	34.24	48.91	36.08	1.5834×10^{1} 3.0323
SUB-SOC-110M-2	25.38	25.93	25.20	51.25	35.57	49.49	35.47	25.26	34.01	48.07	35.77	4.6945×10^{1} 1.4708×10^{1}
ABL-SOC-110M-1	33.00	26.88	27.60	55.50	40.79	51.30	39.18	22.44	36.79	44.66	37.77	1.8421×10^{2} 1.0042×10^{1}
ABL-SOC-110M-2	24.87	26.03	26.60	51.85	36.80	50.04	36.03	24.91	34.44	50.41	36.44	4.4422×10^{1} 5.0448
ABL-SOC-110M-3	25.46	25.45	28.20	52.45	36.54	49.80	36.32	28.16	35.15	48.42	36.81	1.8211×10^{1} 3.7942×10^{3}
GPT-Neo-125M	39.39	30.40	26.20	62.46	42.07	50.91	41.91	23.12	39.22	45.58	40.02	NA

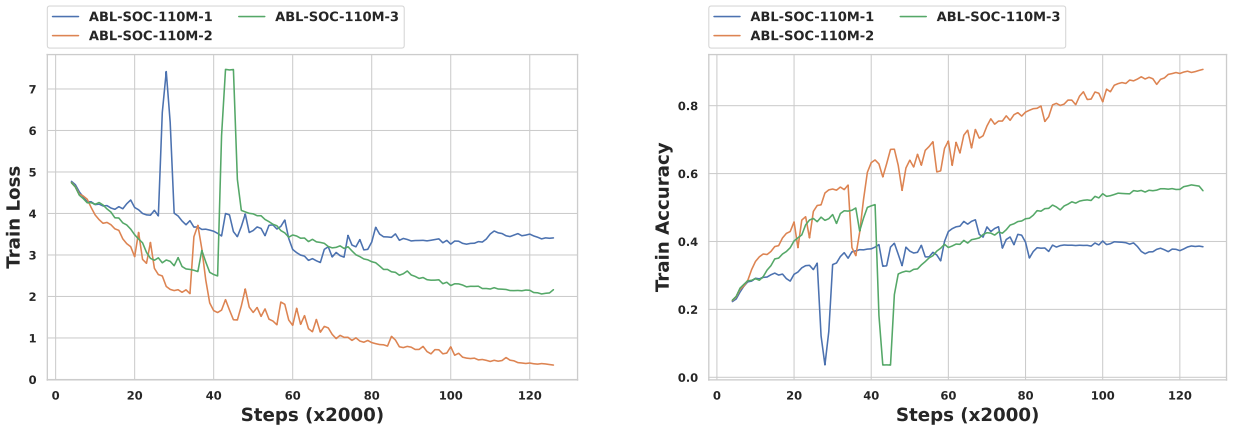
by large normalized RMSE by mean magnitude values in table 6. This observation is also supported by reduced benchmark scores in table 7.

Dragon kings are catastrophic events that appear in critical physical systems in nature and could be very informative of the course of progress in such systems. For PLDR-LLM and other critical systems, dragon kings are predictable events and can be avoided. The histogram distributions of deductive outputs and additional statistics of the ablation models are provided in the appendix.

6 Discussion

The observation that PLDR-LLM attains reasoning at self-organized criticality and the deductive outputs establish an input independent metastable steady state can help to explain several characteristics of LLMs in general. The improvement of LLM reasoning with scaling of model size and training data [Kaplan et al., 2020, Hoffmann et al., 2022] also increases the number of parameters for the steady state which can capture more details of the higher dimensional symmetries of the pretraining data. Since steady state of deductive outputs is unchanged for any input in a model with good reasoning capabilities, scaling the model size becomes a very effective way to improve the reasoning capabilities. It is also easier to understand under the steady state condition, why models with deeper layers often outperform the wider models. The possible combinations of interactions between hidden states increases exponentially in deeper models, which improve the representation capacity of a system with fixed entries at steady state.

The self-organized criticality needs slowly driven updates with infinitesimal length to achieve and maintain the steady state at criticality. Therefore, slower changes of parameters are favored to maintain the delicate balance forming the steady state in each PLGA layer. For example, SwiGLU [Dauphin et al., 2017, Shazeer, 2020] allows a linear path for gradients while maintaining non-linearity and rotary positional embedding [Su et al., 2021] rotate the embedding vectors instead of changing their magnitude by addition of a positional embedding value. We expect that improvements to PLDR-LLM that help maintain the steady state at criticality easier can help develop better performing and more robust language models.



(a) (b)

Figure 3: Train loss (a) and accuracy (b) curves for the PLDR-LLMs pretrained as ablation study near-critical and sub-critical conditions and exhibiting dragon king events. Each data point was captured as a running average of 2000 steps.

PLDR-LLM and its connection to self-organized criticality provide significant insights into how reasoning can arise from a language model. Self-organized criticality is observed in many physical systems, some of which have rich, reliable data to observe and others are very scarce in data. Moreover, these processes may belong to same universality classes. One exciting area of research on PLDR-LLM's capacity to generalize at criticality can be whether a model trained on a high resource field such as NLP can improve prediction capability on a field that is scarce of observations in a controlled environment such as understanding earthquake dynamics. In traditional machine learning, this is possible to a degree with methods such as fine-tuning or transfer learning that improves prediction of the inductive output. However, the model is still a black box with little understanding on how the low resource system is generalized. Access to the details of steady state dynamics through deductive outputs of PLDR-LLMs can further improve our understanding in complex physical systems that are otherwise hard to observe and thus generalize.

One interesting observation is that the steady state condition only perturbs negligibly under stochastic sampling conditions such as nucleus sampling used in this study. In SDPA-LLMs which do not have access to any of the deductive output dynamics, the steady state is a hidden variable (unknown to the observer) and the nucleus sampling of the inductive output is probabilistic. Query and key vectors define attention, fully-connected layers transform the attention at each layer and a choice is made for the next token based on the logits of the probabilities in the final layer. Unlike an SDPA-LLM, deductive outputs of PLDR-LLM at criticality do not get modified with any choice for the next token. This situation suggests that the deductive outputs learn the representations of scaling functions, universality classes and renormalization groups such that for any input it is unchanged. This mechanism of generalization is fundamentally different than how the model learns for its fully-connected layers with activation functions.

It is hypothesized that brain also operates at criticality. Some evidence has been shown to support this hypothesis, though more experiments are needed. As a computational model that is a fully controlled environment, PLDR-LLM can be an important tool to understand the dynamics of human brain further. Possible applications of PLDR-LLM as a laboratory test vehicle can be to understand and treat cognitive disorders, and uncover the differences in reasoning by LLMs compared to the brain.

7 Conclusion

We extended the investigations into unique characteristics of PLDR-LLM architecture observed since its inception from the perspective of self-organized criticality. We show that PLDR-LLM exhibits behaviour similar to second-order phase transitions during training. The reasoning capability is achieved when PLDR-LLMs are pretrained at criticality. The deductive outputs of PLDR-LLM attain a metastable steady-state at inference, and this suggests that PLGA learns the representations equivalent to scaling functions, universality classes and renormalization groups from the pretraining data. This type of generalization through a steady state condition allows the reasoning capability of the model to be quantified very precisely. We define an order parameter from the normalized RMSE by mean magnitude of global deductive output values from separate runs of generic input prompts with stochastic sampling. We show

that this order parameter approaches closer to zero for models that also show higher benchmark scores for reasoning and comprehension. Our results indicate that PLDR-LLM is a self-contained model whose characteristics can be completely determined from the deductive outputs. This observation opens the path for PLDR-LLMs to be studied in every aspect without the need for training very large size models that require extensive computing resources for training and inference. The curated benchmark datasets are insightful for evaluating inductive output of a model, but not needed to quantify reasoning in the self-organized criticality perspective. We believe our understanding of PLDR-LLM and its reasoning dynamics presented here will lead to better analytical characterization of LLMs in general, the ability to better understand complex physical systems in other domains with low resources for observation, and how reasoning manifests in the brain.

Acknowledgments

I am grateful to my parents for their support and patience. This research was conducted independently without support from a grant or corporation.

References

Burc Gokden. Pldr-llms learn a generalizable tensor operator that can replace its own deep neural net at inference, 2025. URL https://arxiv.org/abs/2502.13502.

Burc Gokden. Pldr-llm: Large language model from power law decoder representations, 2024. URL https://arxiv.org/abs/2410.16703.

Burc Gokden. Power law graph transformer for machine translation and representation learning, 2021. URL https://arxiv.org/abs/2107.02039.

Burc Gokden. Coulgat: An experiment on interpretability of graph attention networks, 2019. URL https://arxiv.org/abs/1912.08409.

Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality. *Phys. Rev. A*, 38:364–374, Jul 1988. doi:10.1103/PhysRevA.38.364.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

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. URL https://arxiv.org/abs/2302.13971.

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. URL https://arxiv.org/abs/2307.09288.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 933–941. JMLR.org, 2017.

Dimitrije Marković and Claudius Gros. Power laws and self-organized criticality in theory and nature. *Physics Reports*, 536(2):41–74, 2014. ISSN 0370-1573. doi:https://doi.org/10.1016/j.physrep.2013.11.002.

Markus J. Aschwanden and Ersin Göğüş. Testing the universality of self-organized criticality in galactic, extragalactic, and black hole systems. *The Astrophysical Journal*, 978(1):19, dec 2024. doi:10.3847/1538-4357/ad8dca.

Daniele Notarmuzi, Claudio Castellano, Alessandro Flammini, Dario Mazzilli, and Filippo Radicchi. Universality, criticality and complexity of information propagation in social media. *Nature Communications*, 13:1308, 2022. doi:10.1038/s41467-022-28964-8.

George K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949.

Vasilii A. Gromov and Anastasia M. Migrina. A language as a self-organized critical system. *Complexity*, 2017, January 2017. ISSN 1076-2787. doi:10.1155/2017/9212538.

H. Stanley. Scaling, universality and renormalization: Three pillars of modern critical phenomena. *Rev. Mod. Phys.*, 71:S358, 03 1999. doi:10.1103/RevModPhys.71.S358.

Ronald Dickman, Miguel Angel Muñoz, Alessandro Vespignani, and Stefano Zapperi. Paths to self-organized criticality. *Brazilian Journal of Physics*, 30:27–41, 1999.

John M. Beggs and Dietmar Plenz. Neuronal avalanches in neocortical circuits. *The Journal of Neuroscience*, 23: 11167 – 11177, 2003.

Janina Hesse and Thilo Gross. Self-organized criticality as a fundamental property of neural systems. *Frontiers in Systems Neuroscience*, Volume 8 - 2014, 2014. ISSN 1662-5137. doi:10.3389/fnsys.2014.00166.

Thomas Petermann, Tara C. Thiagarajan, Mikhail A. Lebedev, Miguel A. L. Nicolelis, Dante R. Chialvo, and Dietmar Plenz. Spontaneous cortical activity in awake monkeys composed of neuronal avalanches. *Proceedings of the National Academy of Sciences*, 106(37):15921–15926, 2009. doi:10.1073/pnas.0904089106.

Dietmar Plenz, Tiago L. Ribeiro, Stephanie R. Miller, Patrick A. Kells, Ali Vakili, and Elliott L. Capek. Self-organized criticality in the brain. *Frontiers in Physics*, Volume 9 - 2021, 2021. ISSN 2296-424X. doi:10.3389/fphy.2021.639389. URL `https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2021.639389`.

Tiago L. Ribeiro, Mauro Copelli, Fábio Caixeta, Hindiael Belchior, Dante R. Chialvo, Miguel A. L. Nicolelis, and Sidarta Ribeiro. Spike avalanches exhibit universal dynamics across the sleep-wake cycle. *PLOS ONE*, 5(11):1–14, 11 2010. doi:10.1371/journal.pone.0014129. URL `https://doi.org/10.1371/journal.pone.0014129`.

Yuchen Xie. Bio-inspired llms forgetting: Integrating neuroscience and computational mechanisms. AIVRID '25, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400718540. doi:10.1145/3777730.3777756.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data only. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi:10.18653/v1/D18-2012.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1007.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

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. URL `https://arxiv.org/abs/1803.05457`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi:10.1145/3474381.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.229.

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.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1454.

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. A framework for few-shot language model evaluation, 07 2024. URL `https://zenodo.org/records/12608602`.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL `https://doi.org/10.5281/zenodo.5297715`.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. ICML'13, page III–1310–III–1318. JMLR.org, 2013.

Guram Mikaberidze, Arthur Plaud, and Raissa M. D'Souza. Dragon kings in self-organized criticality systems. *Phys. Rev. Res.*, 5:L042013, Oct 2023. doi:10.1103/PhysRevResearch.5.L042013.

Didier Sornette and Guy Ouillon. Dragon-kings: Mechanisms, statistical methods and empirical evidence. *The European Physical Journal Special Topics*, 205:1 – 26, 2012.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 2022. ISBN 9781713871088.

Noam Shazeer. Glu variants improve transformer, 2020. URL `https://arxiv.org/abs/2002.05202`.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. 2021. URL `https://arxiv.org/abs/2104.09864`.

Appendix

A Mean and Std Dev for Pretrained PLDR-LLM Deductive Outputs

Table 8: Mean (μ) values of $\mathbf{A}$ and $\mathbf{A}_{LM}$ of all pretrained models at runs 1, 2 and Cached.

		$\mathbf{A}$			$\mathbf{A}_{LM}$		
		μ_1	μ_2	μ_{Cached}	μ_1	μ_2	μ_{Cached}
NEAR CRITICAL	PLDRv51-SOC-110M-1	2.1041×10^{-3}	2.1041×10^{-3}	2.1041×10^{-3}	6.0489×10^{-4}	6.0489×10^{-4}	6.0489×10^{-4}
	PLDRv51-SOC-110M-2	-9.8446×10^{-4}	-9.8446×10^{-4}	-9.8446×10^{-4}	1.6681×10^{-4}	1.6681×10^{-4}	1.6681×10^{-4}
	PLDRv51-SOC-110M-3	5.9329×10^{-3}	5.9329×10^{-3}	5.9329×10^{-3}	2.4056×10^{-4}	2.4056×10^{-4}	2.4056×10^{-4}
	PLDRv51-SOC-110M-4	-1.6349×10^{-3}	-1.6349×10^{-3}	-1.6349×10^{-3}	1.6954×10^{-4}	1.6954×10^{-4}	1.6954×10^{-4}
	PLDRv51-SOC-110M-5	-1.3555×10^{-2}	-1.3555×10^{-2}	-1.3555×10^{-2}	2.1645×10^{-2}	2.1645×10^{-2}	2.1645×10^{-2}
SUB CRITICAL	SUB-SOC-110M-1	9.7475×10^{-3}	9.7477×10^{-3}	9.6593×10^{-3}	1.0627×10^{-3}	1.0635×10^{-3}	1.0392×10^{-3}
	SUB-SOC-110M-2	-6.6022×10^{-3}	-6.6019×10^{-3}	-6.9380×10^{-3}	5.4364×10^{-4}	5.4439×10^{-4}	5.1420×10^{-4}
ABLATION	ABL-SOC-110M-1	-1.0951×10^{-2}	-1.0947×10^{-2}	-1.1575×10^{-2}	7.9532×10^{-3}	7.9589×10^{-3}	8.0124×10^{-3}
	ABL-SOC-110M-2	-1.3195×10^{-2}	-1.3183×10^{-2}	-1.3912×10^{-2}	3.9689×10^{-3}	3.9656×10^{-3}	3.9123×10^{-3}
	ABL-SOC-110M-3	1.5216×10^{-4}	1.5748×10^{-4}	1.8376×10^{-5}	3.3616×10^{-4}	3.3625×10^{-4}	3.3062×10^{-4}

Table 9: Mean (μ) values of $\mathbf{A_P}$ and $\mathbf{G}_{LM}$ of all pretrained models at runs 1, 2 and Cached.

		$\mathbf{A_P}$			$\mathbf{G}_{LM}$		
		μ_1	μ_2	μ_{Cached}	μ_1	μ_2	μ_{Cached}
NEAR CRITICAL	PLDRv51-SOC-110M-1	1.2127	1.2127	1.2127	-9.2139×10^{-3}	-9.2140×10^{-3}	-9.2138×10^{-3}
	PLDRv51-SOC-110M-2	1.2592	1.2592	1.2592	-4.9210×10^{-3}	-4.9210×10^{-3}	-4.9210×10^{-3}
	PLDRv51-SOC-110M-3	1.1864	1.1864	1.1864	3.9397×10^{-3}	3.9397×10^{-3}	3.9397×10^{-3}
	PLDRv51-SOC-110M-4	1.2834	1.2834	1.2834	-2.3352×10^{-3}	-2.3352×10^{-3}	-2.3352×10^{-3}
	PLDRv51-SOC-110M-5	1.5005	1.5005	1.5005	-9.3573×10^{-3}	-9.3573×10^{-3}	-9.3573×10^{-3}
SUB CRITICAL	SUB-SOC-110M-1	1.1717	1.1717	1.1719	1.1005×10^{-2}	1.1015×10^{-2}	1.0833×10^{-2}
	SUB-SOC-110M-2	1.1240	1.1237	1.1224	-1.8509×10^{-2}	-1.8685×10^{-2}	-1.9625×10^{-2}
ABLATION	ABL-SOC-110M-1	3.9976	3.9919	4.0119	-4.8490×10^{-2}	-4.8331×10^{-2}	-4.6665×10^{-2}
	ABL-SOC-110M-2	1.1032	1.1031	1.1042	-1.3235×10^{-2}	-1.3150×10^{-2}	-1.2866×10^{-2}
	ABL-SOC-110M-3	1.2135	1.2135	1.2153	-2.2128×10^{-2}	-2.2210×10^{-2}	-2.2251×10^{-2}

Table 10: Standard deviation (σ) values of $\mathbf{A}$ and $\mathbf{A}_{LM}$ of all pretrained models at runs 1, 2 and Cached.

		$\mathbf{A}$			$\mathbf{A}_{LM}$		
		σ_1	σ_2	σ_{Cached}	σ_1	σ_2	σ_{Cached}
NEAR CRITICAL	PLDRv51-SOC-110M-1	5.7461×10^{-2}	5.7461×10^{-2}	5.7461×10^{-2}	4.7671×10^{-3}	4.7671×10^{-3}	4.7671×10^{-3}
	PLDRv51-SOC-110M-2	8.1209×10^{-2}	8.1209×10^{-2}	8.1209×10^{-2}	1.4358×10^{-3}	1.4358×10^{-3}	1.4358×10^{-3}
	PLDRv51-SOC-110M-3	1.0658×10^{-1}	1.0658×10^{-1}	1.0658×10^{-1}	1.7124×10^{-3}	1.7124×10^{-3}	1.7124×10^{-3}
	PLDRv51-SOC-110M-4	9.2655×10^{-2}	9.2655×10^{-2}	9.2655×10^{-2}	6.8940×10^{-4}	6.8940×10^{-4}	6.8940×10^{-4}
	PLDRv51-SOC-110M-5	2.4249×10^{-1}	2.4249×10^{-1}	2.4249×10^{-1}	1.1569×10^{1}	1.1569×10^{1}	1.1569×10^{1}
SUB CRITICAL	SUB-SOC-110M-1	7.0436×10^{-2}	7.0462×10^{-2}	6.9683×10^{-2}	1.0331×10^{-2}	1.0328×10^{-2}	1.0204×10^{-2}
	SUB-SOC-110M-2	1.5189×10^{-1}	1.5190×10^{-1}	1.5245×10^{-1}	6.7867×10^{-3}	6.7761×10^{-3}	6.9008×10^{-3}
ABLATION	ABL-SOC-110M-1	2.2558×10^{-1}	2.2574×10^{-1}	2.1577×10^{-1}	8.1948×10^{-1}	8.2113×10^{-1}	8.1877×10^{-1}
	ABL-SOC-110M-2	1.4103×10^{-1}	1.4106×10^{-1}	1.4308×10^{-1}	3.6208×10^{-2}	3.6157×10^{-2}	3.6568×10^{-2}
	ABL-SOC-110M-3	8.8414×10^{-2}	8.8436×10^{-2}	8.9070×10^{-2}	3.3533×10^{-3}	3.3539×10^{-3}	3.3990×10^{-3}

Table 11: Standard deviation (σ) values of $\mathbf{A_P}$ and $\mathbf{G}_{LM}$ of all pretrained models at runs 1, 2 and Cached.

		$\mathbf{A_P}$			$\mathbf{G}_{LM}$		
		σ_1	σ_2	σ_{Cached}	σ_1	σ_2	σ_{Cached}
NEAR CRITICAL	PLDRv51-SOC-110M-1	2.3447	2.3447	2.3447	1.9497	1.9497	1.9497
	PLDRv51-SOC-110M-2	2.6758	2.6758	2.6758	2.2230	2.2230	2.2230
	PLDRv51-SOC-110M-3	2.0695	2.0695	2.0695	1.8114	1.8114	1.8114
	PLDRv51-SOC-110M-4	2.5361	2.5361	2.5361	2.2214	2.2214	2.2214
	PLDRv51-SOC-110M-5	5.8997	5.8997	5.8997	3.3020	3.3020	3.3020
SUB CRITICAL	SUB-SOC-110M-1	2.0562	2.0516	2.0576	1.9203	1.9190	1.9200
	SUB-SOC-110M-2	2.6259	2.5036	2.6059	1.9082	1.8766	1.8974
ABLATION	ABL-SOC-110M-1	1.4098×10^{2}	1.4015×10^{2}	1.4443×10^{2}	3.6662×10^{1}	3.6502×10^{1}	3.7342×10^{1}
	ABL-SOC-110M-2	1.6127	1.5825	1.9342	1.8571	1.8524	1.9366
	ABL-SOC-110M-3	2.7258	2.7307	2.8533	2.2626	2.2646	2.3034

B RMSE and Normalized RMSE for Pretrained PLDR-LLM Deductive Outputs

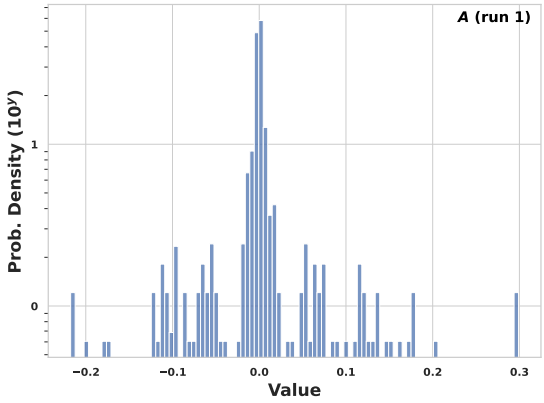
Table 12: RMSE values of deductive outputs of all pretrained models at runs 1, 2 and Cached.

		$\mathbf{A}$		$\mathbf{A}_{LM}$		$\mathbf{A}_{\mathbf{P}}$		$\mathbf{G}_{LM}$	
		$RMSE_{12}$	$RMSE_{1C}$	$RMSE_{12}$	$RMSE_{1C}$	$RMSE_{12}$	$RMSE_{1C}$	$RMSE_{12}$	$RMSE_{1C}$
NEAR CRITICAL	PLDRv51-SOC-110M-1	1.7823×10^{-5}	1.7479×10^{-5}	4.5307×10^{-7}	4.1603×10^{-7}	7.9668×10^{-4}	1.1559×10^{-3}	1.6772×10^{-4}	2.0148×10^{-4}
	PLDRv51-SOC-110M-2	4.1713×10^{-9}	4.2848×10^{-9}	1.3083×10^{-10}	1.4341×10^{-10}	1.2817×10^{-7}	1.2633×10^{-7}	6.2531×10^{-8}	6.1651×10^{-8}
	PLDRv51-SOC-110M-3	8.1415×10^{-7}	1.0519×10^{-6}	1.5399×10^{-10}	1.9643×10^{-10}	3.4293×10^{-6}	3.8472×10^{-6}	1.8074×10^{-6}	2.0227×10^{-6}
	PLDRv51-SOC-110M-4	2.1092×10^{-9}	2.1410×10^{-9}	3.0347×10^{-11}	3.1737×10^{-11}	4.7302×10^{-8}	4.6144×10^{-8}	2.1934×10^{-8}	2.1485×10^{-8}
	PLDRv51-SOC-110M-5	4.7764×10^{-13}	5.6471×10^{-13}	2.3071×10^{-14}	2.7074×10^{-14}	0	0	0	0
SUB CRITICAL	SUB-SOC-110M-1	2.8131×10^{-2}	2.9290×10^{-2}	3.8573×10^{-3}	3.8069×10^{-3}	2.1001×10^{-1}	3.9297×10^{-1}	9.6239×10^{-2}	1.7153×10^{-1}
	SUB-SOC-110M-2	4.7210×10^{-2}	1.0205×10^{-1}	1.7820×10^{-3}	3.9981×10^{-3}	1.5788	2.2192	6.4775×10^{-1}	9.2130×10^{-1}
ABLATION	ABL-SOC-110M-1	8.5889×10^{-2}	1.1624×10^{-1}	9.8798×10^{-2}	1.1974×10^{-1}	2.4309×10^{1}	3.4371×10^{1}	6.4290	8.5961
	ABL-SOC-110M-2	5.9923×10^{-2}	7.0185×10^{-2}	1.1403×10^{-2}	1.2688×10^{-2}	3.9351×10^{-1}	1.2008	1.6470×10^{-1}	5.7152×10^{-1}
	ABL-SOC-110M-3	3.8996×10^{-2}	6.9723×10^{-2}	1.3239×10^{-3}	2.5769×10^{-3}	3.1748×10^{-1}	8.2056×10^{-1}	1.5110×10^{-1}	4.0520×10^{-1}

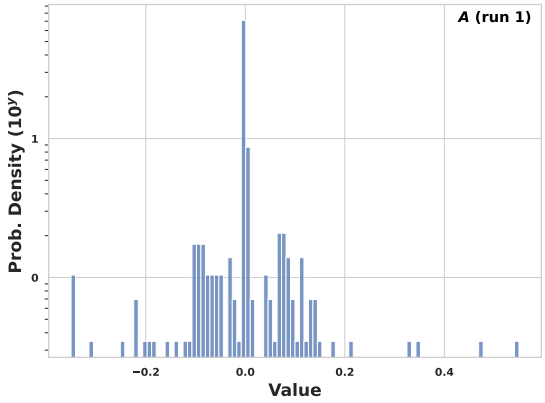
Table 13: Normalized RMSE by mean magnitude of deductive outputs of all pretrained models at runs 1, 2 and Cached.

		$\mathbf{A}$		$\mathbf{A}_{LM}$		$\mathbf{A}_{\mathbf{P}}$		$\mathbf{G}_{LM}$	
		$RMSE_{12}/\lvert\mu_1\rvert$	$RMSE_{1C}/\lvert\mu_C\rvert$	$RMSE_{12}/\lvert\mu_1\rvert$	$RMSE_{1C}/\lvert\mu_C\rvert$	$RMSE_{12}/\lvert\mu_1\rvert$	$RMSE_{1C}/\lvert\mu_C\rvert$	$RMSE_{12}/\lvert\mu_1\rvert$	$RMSE_{1C}/\lvert\mu_C\rvert$
NEAR CRITICAL	PLDRv51-SOC-110M-1	8.4709×10^{-3}	8.3069×10^{-3}	7.4902×10^{-4}	6.8777×10^{-4}	6.5694×10^{-4}	9.5313×10^{-4}	1.8203×10^{-2}	2.1867×10^{-2}
	PLDRv51-SOC-110M-2	4.2371×10^{-6}	4.3525×10^{-6}	7.8430×10^{-7}	8.5975×10^{-7}	1.0178×10^{-7}	1.0032×10^{-7}	1.2707×10^{-5}	1.2528×10^{-5}
	PLDRv51-SOC-110M-3	1.3723×10^{-4}	1.7730×10^{-4}	6.4012×10^{-7}	8.1658×10^{-7}	2.8905×10^{-6}	3.2427×10^{-6}	4.5876×10^{-4}	5.1342×10^{-4}
	PLDRv51-SOC-110M-4	1.2901×10^{-6}	1.3096×10^{-6}	1.7899×10^{-7}	1.8719×10^{-7}	3.6856×10^{-8}	3.5953×10^{-8}	9.3926×10^{-6}	9.2006×10^{-6}
	PLDRv51-SOC-110M-5	3.5237×10^{-11}	4.1660×10^{-11}	1.0659×10^{-12}	1.2508×10^{-12}	0	0	0	0
SUB CRITICAL	SUB-SOC-110M-1	2.8859	3.0323	3.6298	3.6632	1.7923×10^{-1}	3.3531×10^{-1}	8.7449	1.5834×10^{1}
	SUB-SOC-110M-2	7.1506	1.4708×10^{1}	3.2778	7.7753	1.4047	1.9771	3.4997×10^{1}	4.6945×10^{1}
ABLATION	ABL-SOC-110M-1	7.8433	1.0042×10^{1}	1.2422×10^{1}	1.4944×10^{1}	6.0810	8.5672	1.3258×10^{2}	1.8421×10^{2}
	ABL-SOC-110M-2	4.5414	5.0448	2.8730	3.2431	3.5674×10^{-1}	1.0874	1.2445×10^{1}	4.4422×10^{1}
	ABL-SOC-110M-3	2.5628×10^{2}	3.7942×10^{3}	3.9382	7.7942	2.6163×10^{-1}	6.7517×10^{-1}	6.8287	1.8211×10^{1}

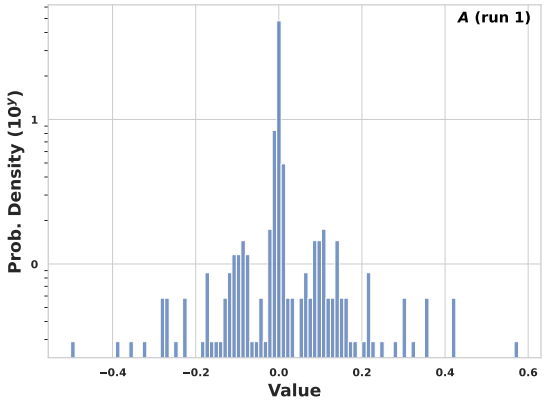
C Global Density Distributions and Heatmaps for Deductive Outputs of PLDR-LLMs



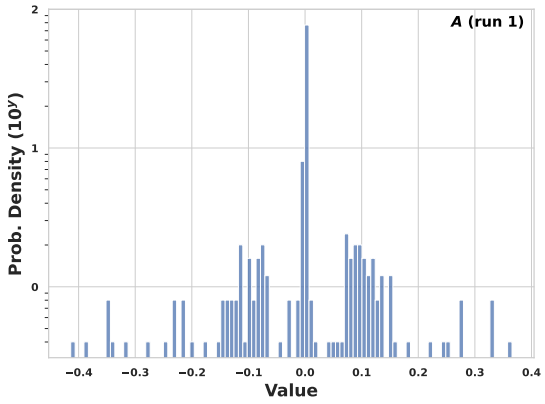
(a) PLDRv51-SOC-110M-1



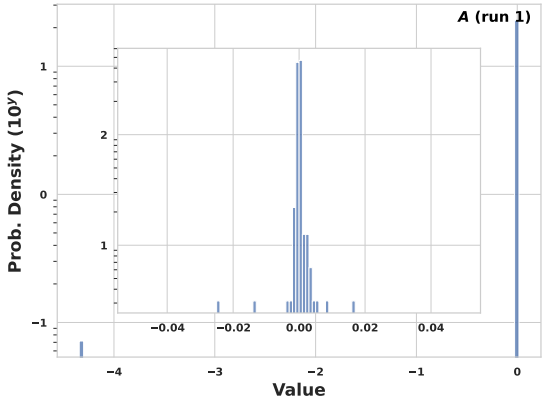
(b) PLDRv51-SOC-110M-2



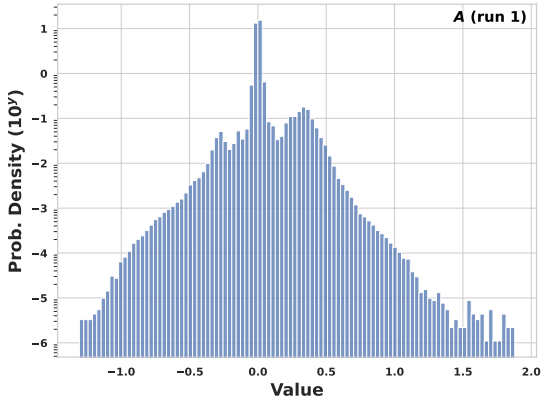
(c) PLDRv51-SOC-110M-3



(d) PLDRv51-SOC-110M-4

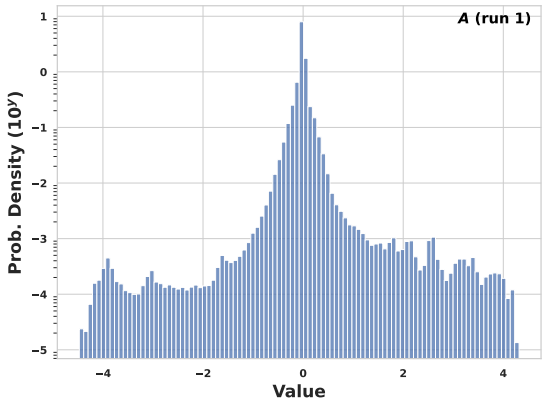


(e) PLDRv51-SOC-110M-5

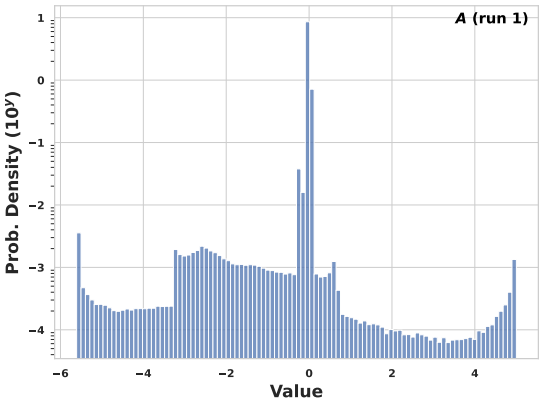


(f) SUB-SOC-110M-1

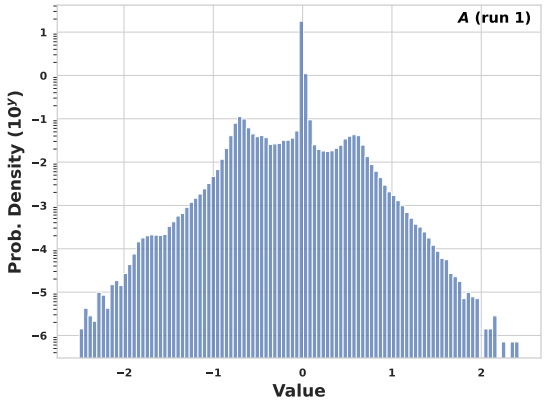
Figure 4: **A** probability density distributions for all models binned in 100 buckets for main plots and insets.



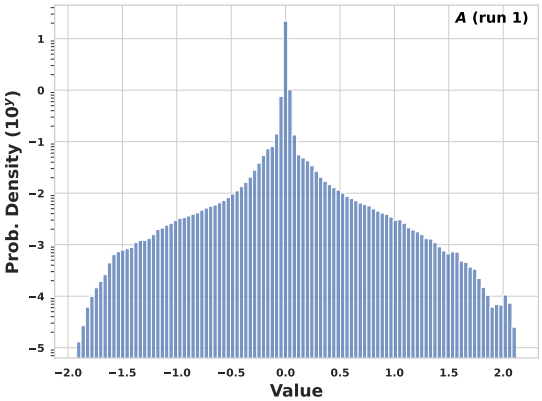
(g) SUB-SOC-110M-2



(h) ABL-SOC-110M-1

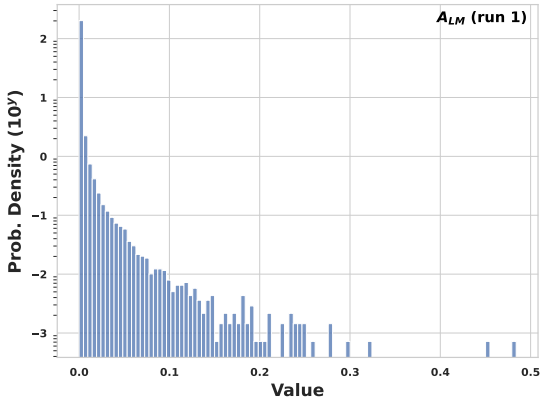


(i) ABL-SOC-110M-2

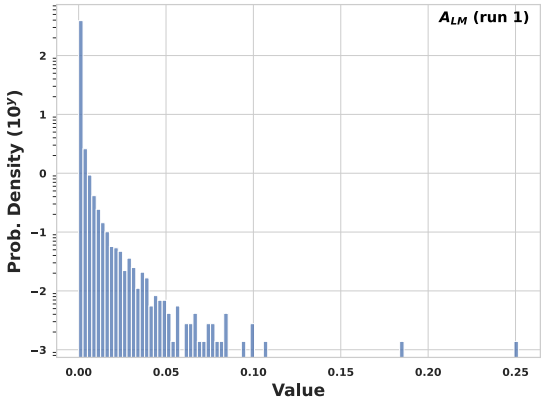


(j) ABL-SOC-110M-3

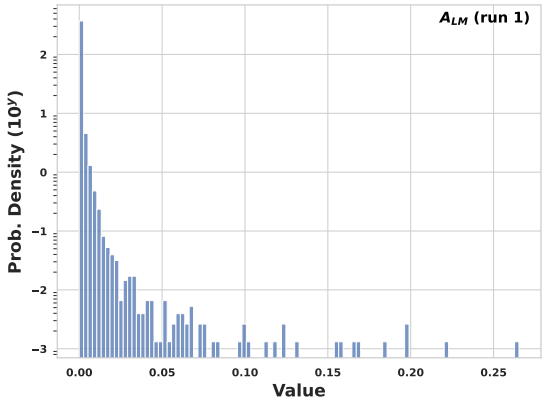
Figure 4: (Cont.) **A** probability density distributions for all models binned in 100 buckets for main plots and insets.



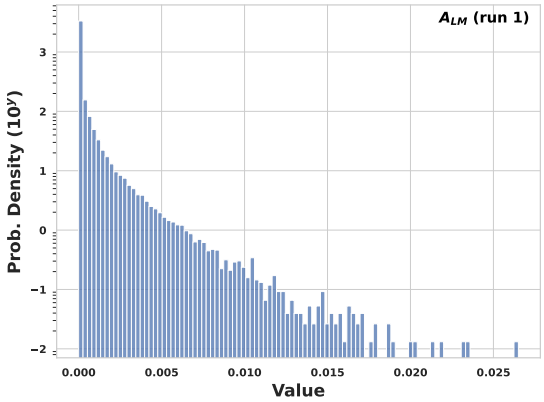
(a) PLDRv51-SOC-110M-1



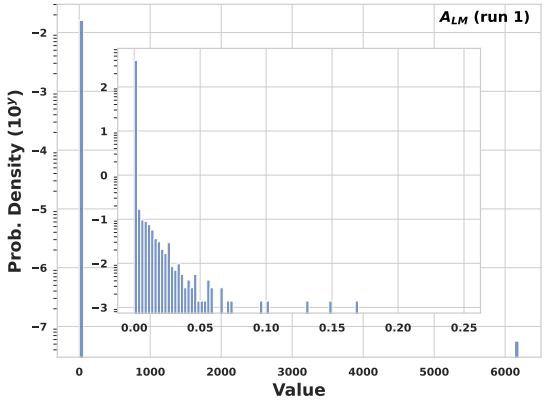
(b) PLDRv51-SOC-110M-2



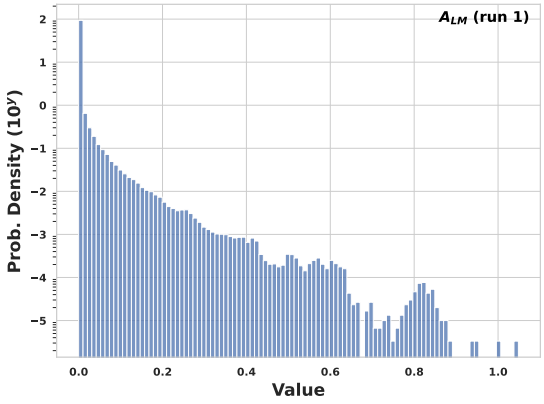
(c) PLDRv51-SOC-110M-3



(d) PLDRv51-SOC-110M-4

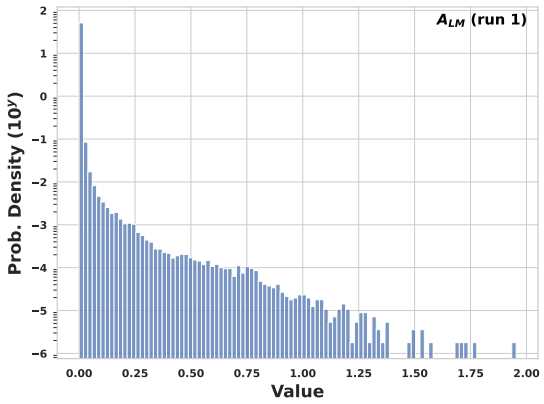


(e) PLDRv51-SOC-110M-5

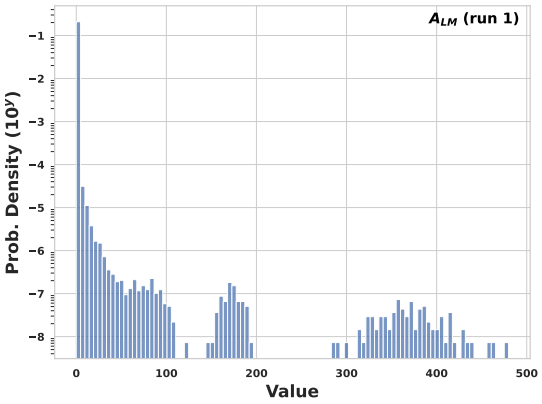


(f) SUB-SOC-110M-1

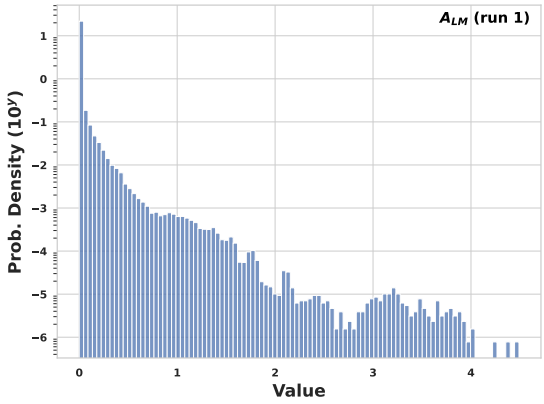
Figure 5: $\mathbf{A}_{LM}$ probability density distributions for all models binned in 100 buckets for main plots and insets.



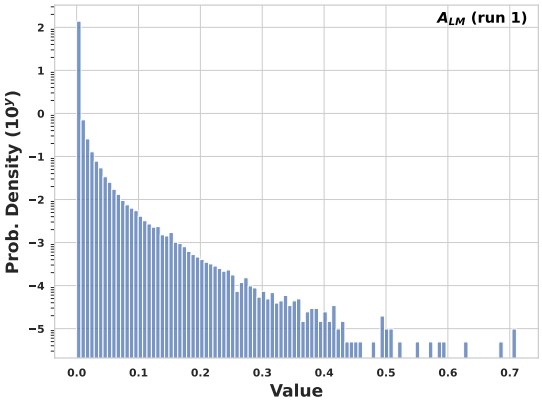
(g) SUB-SOC-110M-2



(h) ABL-SOC-110M-1

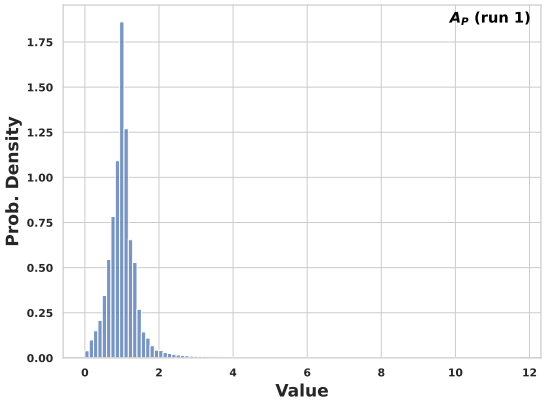


(i) ABL-SOC-110M-2

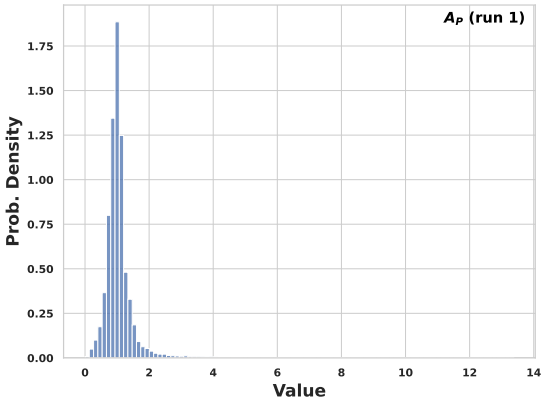


(j) ABL-SOC-110M-3

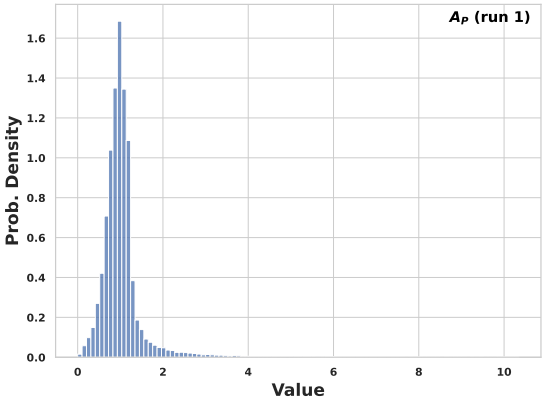
Figure 5: (Cont.) $\mathbf{A}_{LM}$ probability density distributions for all models binned in 100 buckets for main plots and insets.



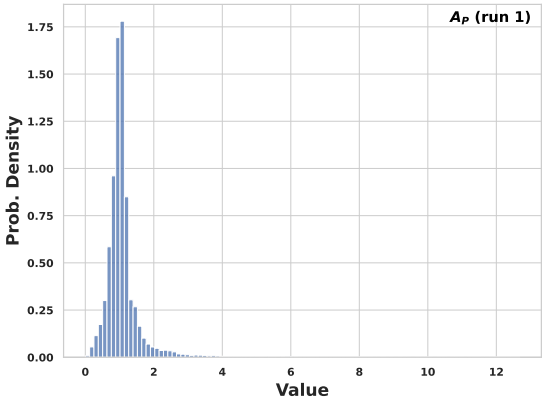
(a) PLDRv51-SOC-110M-1



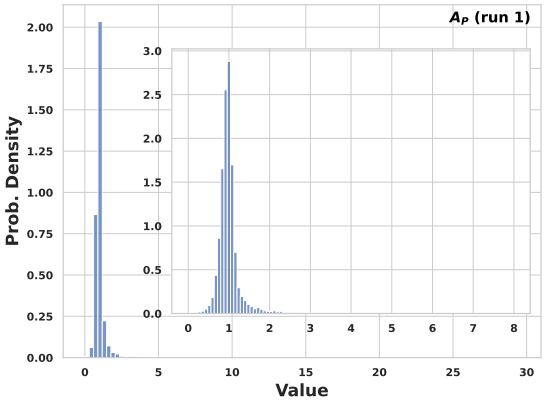
(b) PLDRv51-SOC-110M-2



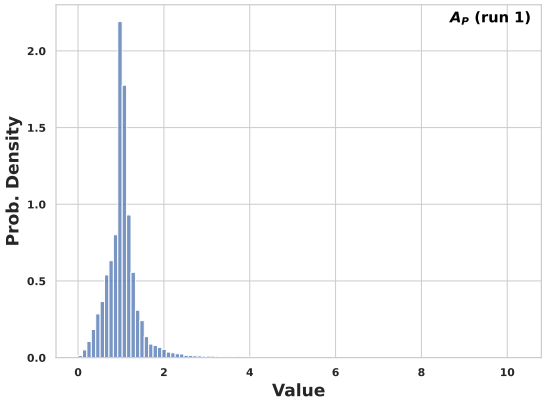
(c) PLDRv51-SOC-110M-3



(d) PLDRv51-SOC-110M-4

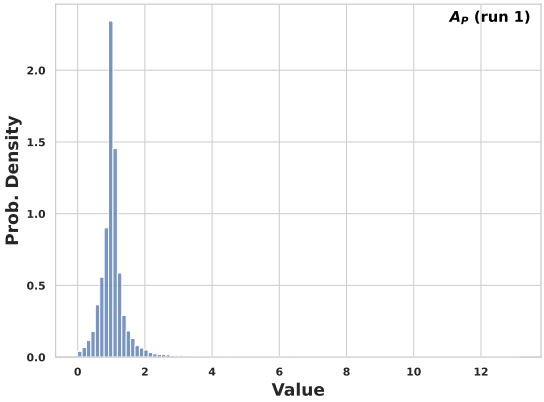


(e) PLDRv51-SOC-110M-5

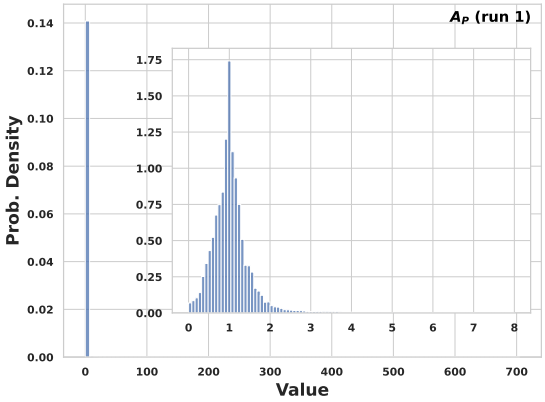


(f) SUB-SOC-110M-1

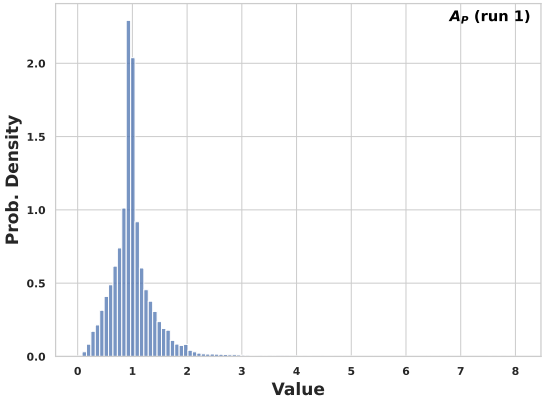
Figure 6: $\mathbf{A_P}$ probability density distributions for all models binned in 100 buckets for main plots and insets. The $\mathbf{A_P}$ were plotted up to $\pm 5\sigma$ for easier visibility of main distribution characteristics.



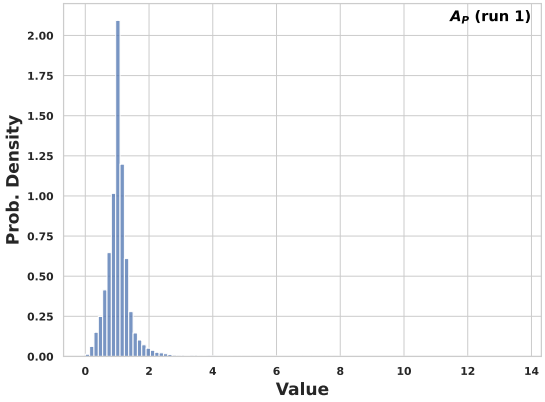
(g) SUB-SOC-110M-2



(h) ABL-SOC-110M-1

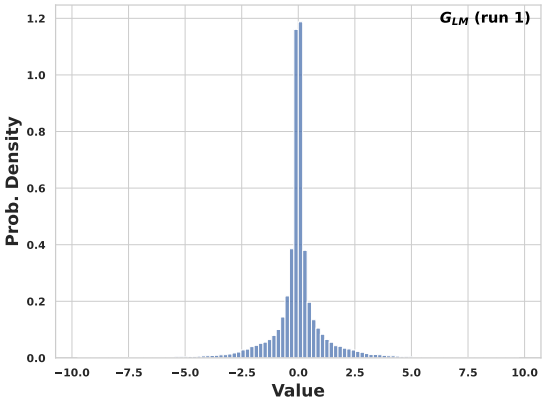


(i) ABL-SOC-110M-2

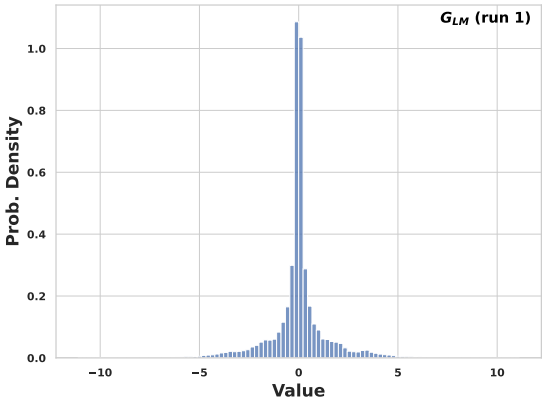


(j) ABL-SOC-110M-3

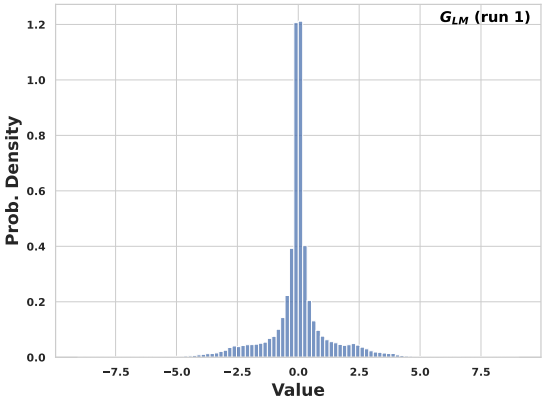
Figure 6: (Cont.) $\mathbf{A_P}$ probability density distributions for all models binned in 100 buckets for main plots and insets. The $\mathbf{A_P}$ were plotted up to $\pm 5\sigma$ for easier visibility of main distribution characteristics.



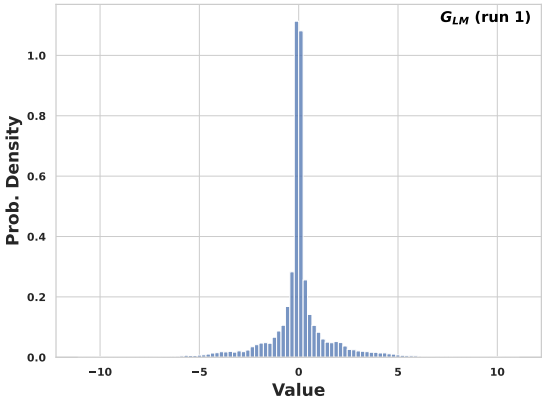
(a) PLDRv51-SOC-110M-1



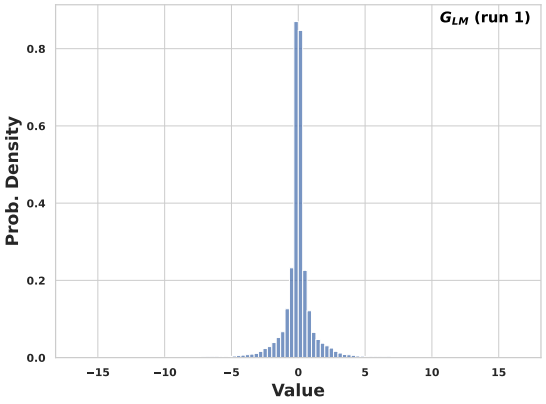
(b) PLDRv51-SOC-110M-2



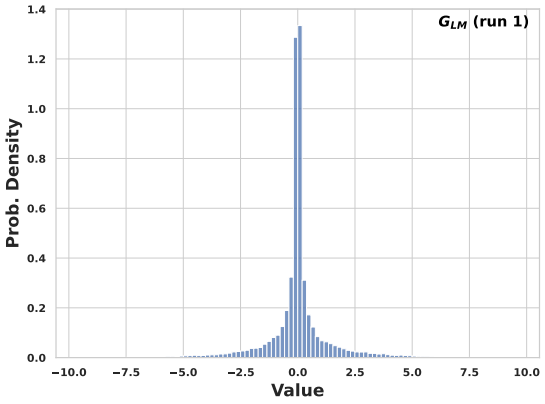
(c) PLDRv51-SOC-110M-3



(d) PLDRv51-SOC-110M-4

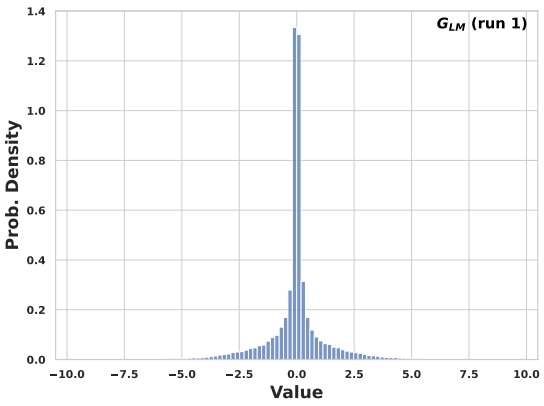


(e) PLDRv51-SOC-110M-5

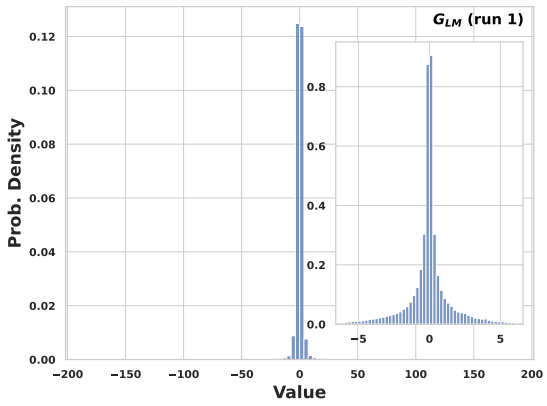


(f) SUB-SOC-110M-1

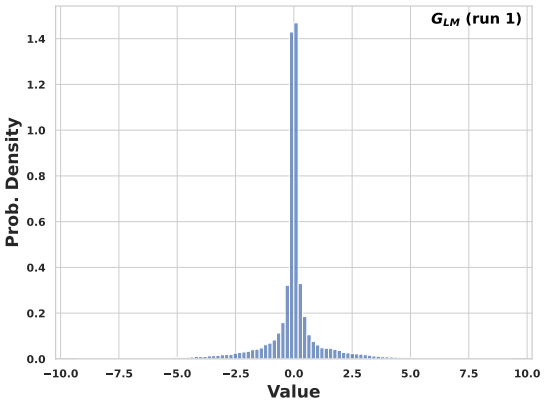
Figure 7: $\mathbf{G}_{LM}$ probability density distributions for all models binned in 100 buckets. The $\mathbf{G}_{LM}$ were plotted up to $\pm 5\sigma$ for easier visibility of main distribution characteristics.



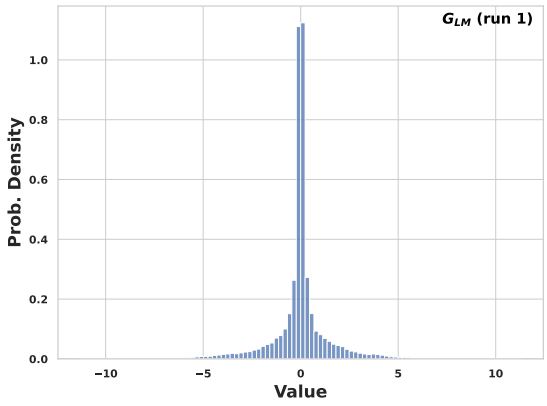
(g) SUB-SOC-110M-2



(h) ABL-SOC-110M-1

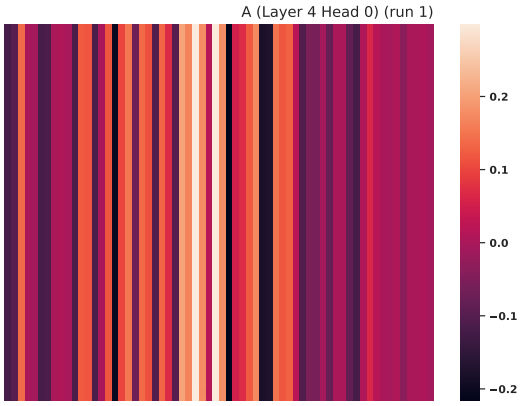


(i) ABL-SOC-110M-2

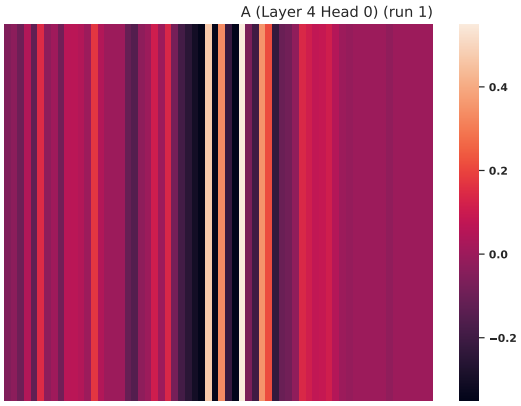


(j) ABL-SOC-110M-3

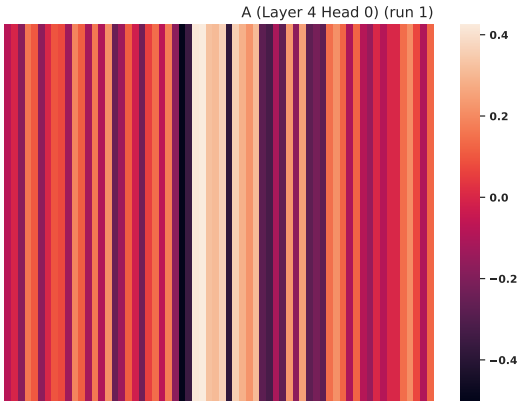
Figure 7: (Cont.) $\mathbf{G}_{LM}$ probability density distributions for all models binned in 100 buckets. Inset for ABL-SOC-110M-1 was binned in 50 buckets. The $\mathbf{G}_{LM}$ were plotted up to $\pm 5\sigma$ for easier visibility of main distribution characteristics.



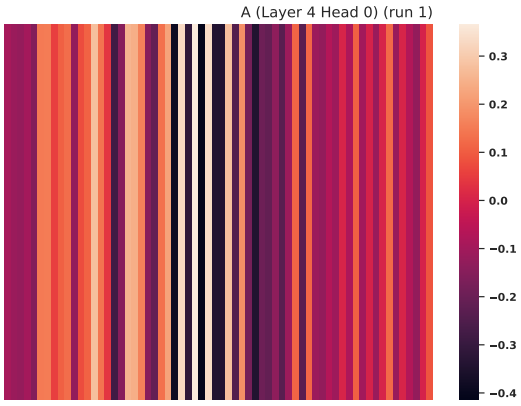
(a) PLDRv51-SOC-110M-1



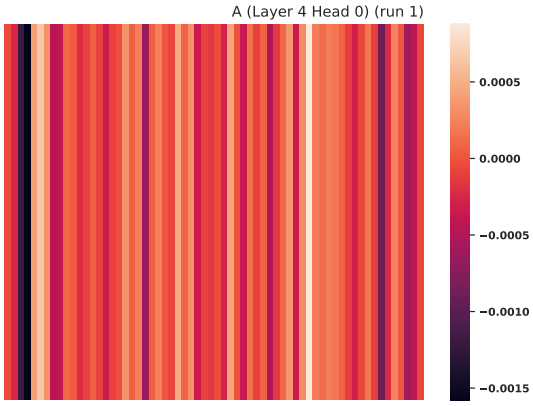
(b) PLDRv51-SOC-110M-2



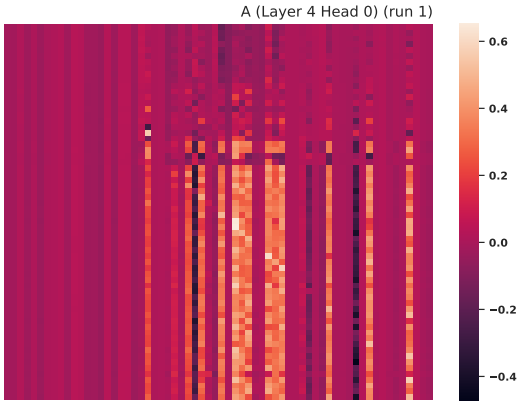
(c) PLDRv51-SOC-110M-3



(d) PLDRv51-SOC-110M-4

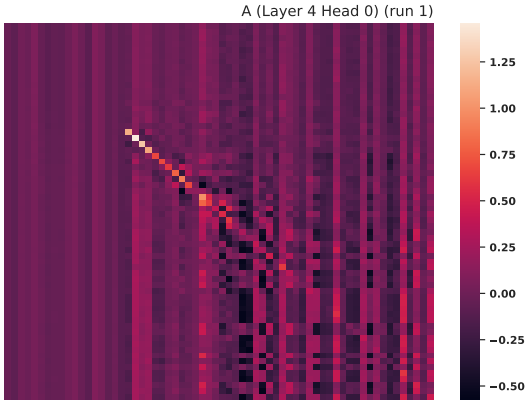


(e) PLDRv51-SOC-110M-5

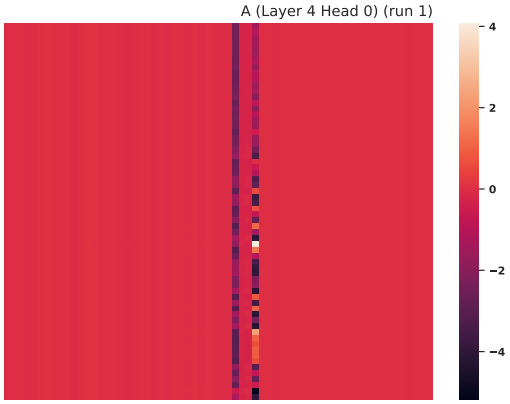


(f) SUB-SOC-110M-1

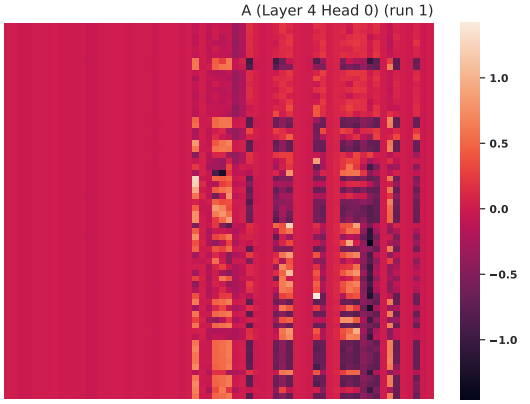
Figure 8: **A** heat maps at last decoder layer and for a single head of all models averaged over all samples.



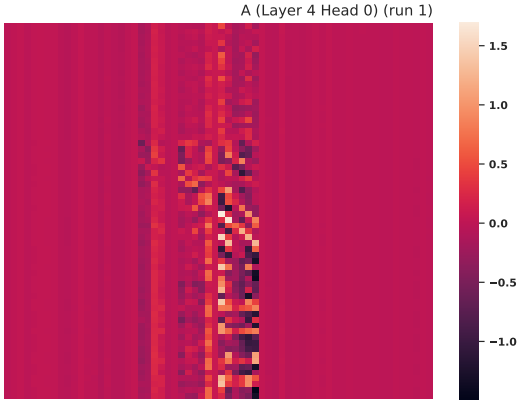
(g) SUB-SOC-110M-2



(h) ABL-SOC-110M-1



(i) ABL-SOC-110M-2



(j) ABL-SOC-110M-3

Figure 8: (Cont.) **A** heat maps at last decoder layer and for a single head of all models averaged over all samples.

D Benchmark Datasets

ARC. The AI2 Reasoning Challenge (ARC) dataset consists of multiple-choice grade school questions from 3^{rd} to 9^{th} grade. It consists of an easy set and a challenge set. The challenge set contains the questions answered incorrectly by both a retrieval based algorithm and a word co-occurrence algorithm [Clark et al., 2018].

Hellaswag. Harder Endings, Longer contexts, and Low-shot Activities for Situations With Adversarial Generations dataset is a commonsense natural language inference dataset that was prepared using adversarial filtering to create problems that are challenging to models, yet easy for humans [Zellers et al., 2019].

WinoGrande. WinoGrande is a more challenging version of Winograd Schema Challenge that is a commonsense reasoning benchmark based on a set of pronoun resolution problems designed to be unsolvable for statistical models that rely on selectional preferences or word associations [Sakaguchi et al., 2021].

TruthfulQA. TruthfulQA is a benchmark that aims to measure truthfullness of a model. It consists of questions covering 38 categories such as health, law, finance and politics. The model should avoid imitating human contexts in pretraining dataset to perform well, since the questions are selected from the ones humans would answer incorrectly due to a false belief or misconception [Lin et al., 2022].

OpenBookQA. OpenBookQA is a question answering dataset that consists of about 6000 questions accompanied with scientific facts. To answer the questions correctly the model needs to combine with extra common knowledge beyond the facts included in the dataset [Mihaylov et al., 2018].

PIQA. Physical Interaction:Question Answering dataset is a physical commonsense benchmark that aims to evaluate model performance for concepts that are traditionally only seen or experienced in the real world [Bisk et al., 2020].

SIQA. Social Intelligence QA dataset is a social commonsense reasoning benchmark that aims to evaluate model performance for social situations. It consists of 38000 multiple-choice questions for probing emotional and social intelligence in a variety of everyday situations [Sap et al., 2019].

IMDB Review. IMDB Review dataset is a collection of 50000 reviews with each movie having no more than 30 reviews. It was compiled for sentiment analysis and consists of an even number of highly polarized negative (≤ 4 out of 10) and positive (≥ 7 out of 10) reviews [Maas et al., 2011].