Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.072735

ARTICLE

Modeling Pruning as a Phase Transition: A Thermodynamic Analysis of Neural Activations

Rayeesa Mehmood*, Sergei Koltcov, Anton Surkov and Vera Ignatenko

Laboratory for Social & Cognitive Informatics, National Research University Higher School of Economics, Sedova St. 55/2, Saint Petersburg, 192148, Russia

*Corresponding Author: Rayeesa Mehmood. Email: rmehmood@hse.ru

Received: 02 September 2025; Accepted: 28 November 2025; Published: 12 January 2026

ABSTRACT: Activation pruning reduces neural network complexity by eliminating low-importance neuron activations, yet identifying the critical pruning threshold—beyond which accuracy rapidly deteriorates—remains computationally expensive and typically requires exhaustive search. We introduce a thermodynamics-inspired framework that treats activation distributions as energy-filtered physical systems and employs the free energy of activations as a principled evaluation metric. Phase-transition–like phenomena in the free-energy profile—such as extrema, inflection points, and curvature changes—yield reliable estimates of the critical pruning threshold, providing a theoretically grounded means of predicting sharp accuracy degradation. To further enhance efficiency, we propose a renormalized free energy technique that approximates full-evaluation free energy using only the activation distribution of the unpruned network. This eliminates repeated forward passes, dramatically reducing computational overhead and achieving speedups of up to 550× for MLPs. Extensive experiments across diverse vision architectures (MLP, CNN, ResNet, MobileNet, Vision Transformer) and text models (LSTM, BERT, ELECTRA, T5, GPT-2) on multiple datasets validate the generality, robustness, and computational efficiency of our approach. Overall, this work establishes a theoretically grounded and practically effective framework for activation pruning, bridging the gap between analytical understanding and efficient deployment of sparse neural networks.

KEYWORDS: Thermodynamics; activation pruning; model compression; sparsity; free energy; renormalization

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success across computer vision, natural language processing, and reinforcement learning. However, their increasing complexity and overparameterization raise growing concerns regarding computational efficiency, energy consumption, and deployment in resource-constrained environments. Consequently, network pruning—the process of reducing model complexity by removing redundant components—has become a key strategy for compressing models while preserving accuracy [1].

Most existing pruning approaches focus on weight magnitudes or structural heuristics, often overlooking the dynamic behavior of activations during inference. In contrast, activation pruning selectively suppresses neuron activations based on their estimated importance, offering a complementary and adaptable mechanism for runtime sparsification [2]. Despite its promise, activation pruning faces a fundamental challenge: reliably identifying the critical pruning threshold, the point at which accuracy begins to degrade sharply. Determining this threshold typically depends on exhaustive trial-and-error procedures, which are

computationally costly and yield inconsistent results across architectures and datasets. The lack of a unified theoretical framework therefore limits both the practicality and widespread adoption of activation pruning.

To address this challenge, we introduce a thermodynamics-inspired framework that interprets activation distributions as physical systems subject to energy filtering. Within this formulation, we employ the free energy of activations as a principled metric for characterizing network behavior under varying pruning thresholds. Salient features of the resulting free-energy curve—such as extrema, inflection points, or curvature changes—serve as indicators of critical pruning thresholds, providing a theoretically grounded alternative to exhaustive empirical search. Moreover, we develop a renormalized free energy that approximates the full free-energy profile using only the activation statistics of the unpruned network ($\tau = 0$). This approach eliminates the need for repeated inference, delivers substantial computational speedups, and maintains high fidelity in estimating optimal pruning thresholds.

Our main contributions are summarized as follows:

- We introduce a thermodynamics-based theoretical framework for activation pruning, in which free energy provides a principled estimator of critical pruning thresholds and corresponding sparsity levels.
- We propose a renormalized free energy approximation that accurately tracks the full-evaluation free energy curve, eliminates repeated inference, and substantially reduces computational cost.
- We conduct an extensive experimental evaluation on both vision and text classification tasks, covering pretrained and non-pretrained models, and demonstrate the efficiency, robustness, and broad applicability of the proposed approach.

The remainder of the paper is organized as follows: Section 2 reviews related work and outlines the motivation for adopting a thermodynamic perspective. Section 3 formalizes the framework, presents the phase transition analysis, and introduces the renormalized free energy method. Section 4 describes the datasets, models, and experimental setup, while Section 5 reports the empirical results. Finally, Section 6 summarizes the main findings, discusses limitations, and outlines directions for future work.

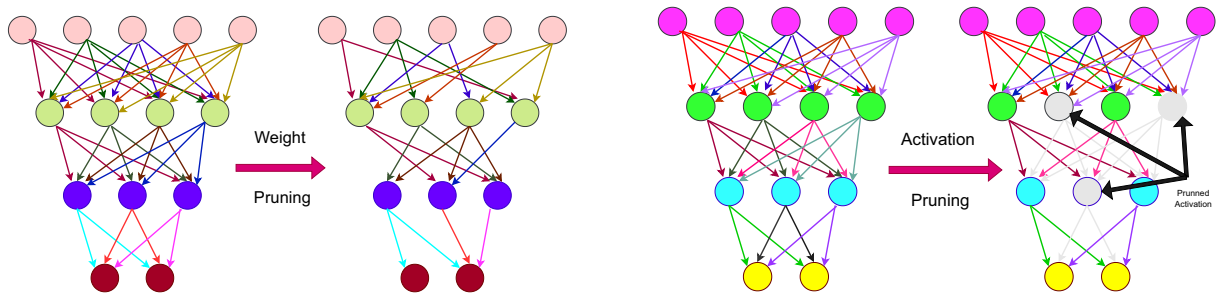
2 Related Work

2.1 Magnitude Pruning: Weight and Activation Pruning

Magnitude pruning is a widely used technique for compressing deep neural networks (DNNs) by removing low-magnitude elements under the assumption that they exert minimal influence on overall performance. Existing methods largely fall into two categories: weight pruning, which operates on static model parameters, and activation pruning, which dynamically sparsifies neuron outputs during inference. Fig. 1 provides a visual comparison of these two approaches.

2.1.1 Weight Pruning

Weight pruning reduces model size and computational cost by removing weights deemed unimportant, typically based on their magnitudes. Han et al. [3] demonstrated that unstructured pruning can compress models such as AlexNet and VGG-16 by more than 90% with minimal loss in accuracy. While unstructured pruning, which removes individual weights, achieves high compression ratios, it yields irregular sparsity patterns that are difficult to exploit efficiently on standard hardware. In contrast, structured pruning removes entire filters, channels, or layers, resulting in more hardware-friendly sparsity. Recent advances incorporate group-wise dependency preservation, gradient-based saliency metrics, and second-order sensitivity analyses to enable more precise and reliable pruning decisions [4–6]. Despite its effectiveness across domains—including reinforcement learning and edge deployment—weight pruning remains a static, input-independent approach.



(a) Weight Pruning: Left—Before pruning. Right—After pruning. Colored edges represent weights; removed weights are absent after pruning.

(b) Activation Pruning: Left—Before pruning. Right—After pruning. Faded neurons indicate pruned (inactive) activations.

Figure 1: Visualization of weight and activation pruning

2.1.2 Activation Pruning

Activation pruning dynamically zeros neuron activations during inference based on a threshold τ, introducing input-dependent sparsity that can substantially reduce computation, particularly on hardware optimized for sparse operations [2]. Early work showed that many neurons become functionally dormant and can be removed using activation statistics [7]. Activation statistics have also been employed to guide structural pruning during training, as in activation-density–based methods [8]. More recent findings [9] reveal that large neural networks exhibit strong intrinsic activation sparsity: transformer layers naturally produce highly sparse activation maps, with only a small fraction of non-zero entries. Enforcing additional sparsity through Top-k thresholding—explicitly zeroing all but the largest k activations—has been shown to improve computational efficiency and, in some cases, enhance robustness.

A variety of activation pruning methods have been proposed, including spatially adaptive sparsification [10], ReLU pruning [11,12], iterative pruning based on activation statistics [13], Wanda pruning [2], attention-guided structured pruning [14], and semi-structured activation sparsity [15]. Recent work [16] introduces WAS, a training-free method that computes weight-aware activation importance and uses constrained Bayesian optimization to allocate sparsity across transformer blocks. Another study [17] proposes R-Sparse, a training-free activation pruning technique for large language models that applies a rank-aware thresholding rule to eliminate low-magnitude activations and reduce unnecessary weight computations. The authors of work [18] present TEAL, a training-free magnitude-based activation pruning method that zeros low-magnitude activations across transformer blocks, achieving substantial sparsity and meaningful inference speedups with limited impact on model quality. Finally, the authors of work [19] introduce Amber Pruner, a training-free N:M activation pruning method that applies structured top-k sparsification to input activations in LLM linear layers during the prefill phase, achieving a reduction of more than 55% of the computation with minimal accuracy loss.

Unlike weight pruning, activation pruning produces dynamic, input-dependent sparsity patterns, which complicates theoretical analysis. Although such techniques leverage activation behavior to improve computational efficiency, they typically depend on empirically tuned thresholds [20], and the underlying theory remains underdeveloped. Concepts such as the activation blind range provide partial insight into zero-output regions [11], yet no unified framework currently exists for predicting critical pruning thresholds or determining optimal sparsity levels.

Beyond classical structured and unstructured pruning approaches, recent years have seen the development of adaptive and one-shot methods capable of achieving high sparsity without retraining [21–23].

These techniques primarily aim to reduce computational cost and accelerate inference. In contrast, our work focuses on analyzing the thermodynamic behavior of neural representations and exploiting the free-energy profile to identify the phase-like transition at which the model begins to lose functional stability. The proposed criterion is therefore not intended to replace existing pruning strategies, but rather to complement them by offering an interpretable, architecture-agnostic means of determining reliable sparsity levels.

2.2 Thermodynamic Perspective on Activation Pruning

Given the heuristic nature of existing activation pruning approaches, a principled theoretical framework is required to estimate when pruning begins to degrade performance and to guide the selection of effective pruning thresholds. To address this need, we introduce a thermodynamics-inspired perspective in which neuron activations are treated analogously to energy states in physical systems.

Within this framework, pruning is interpreted as imposing an energy cutoff, where removing low-activation neurons corresponds to filtering out low-energy states. This analogy enables the definition of thermodynamic quantities such as entropy, internal energy, and free energy over activation distributions, providing a quantitative basis for analyzing how pruning influences model behavior. Notably, we observe that model performance remains stable over a broad range of pruning thresholds but deteriorates sharply beyond a critical point, resembling a phase transition. This transition manifests in distinct features of the free-energy curve, which serve as reliable indicators of critical thresholds without requiring exhaustive empirical tuning.

The proposed formulation offers theoretical insight into the trade-off between pruning intensity and predictive accuracy, while enhancing interpretability and robustness across architectures and datasets. In the following section, we formalize these concepts, introduce precise definitions, and derive the free-energy criterion used to approximate critical pruning thresholds.

3 Formal Thermodynamic Framework and Phase Transition Analysis

Building on the thermodynamic perspective introduced in Section 2.2, we now develop a formal framework that models activation pruning using concepts from statistical physics. Specifically, we define entropy, internal energy, and free energy over activation distributions, and show how these quantities expose critical thresholds at which model behavior undergoes abrupt transitions. This formulation establishes a unified and interpretable connection between sparsity, thresholding, and accuracy degradation.

3.1 Activation Pruning as Energy-Based Filtering

We model the neural network as a statistical system characterized by the distribution of neuron activations during inference. Activation pruning is implemented by applying a threshold τ such that activations a_i with magnitudes below τ are set to zero, effectively filtering out low-energy microstates. Eq. (1) formalizes this operation, where a_i' denotes the post-pruning activation values and $\mathbf{1}_{\{\cdot\}}$ is the indicator function:

$$a_i' = a_i \cdot \mathbf{1}_{\{|a_i| \geq \tau\}}. \tag{1}$$

The pruning process is applied layer-wise, and for each threshold τ we compute two key quantities: (1) the number of nonzero activations M, and (2) the sum of their absolute values Σ. These values are aggregated across all layers to characterize the state of the system. In practice, activations are collected selectively from nonlinear and fully connected layers, as these dominate the activation distribution relevant to pruning.

3.2 Shannon Entropy of Activations

Let M denote the number of nonzero activations after pruning and N the total number of activations before pruning (i.e., at $\tau = 0$). We define the density of nonzero activations as $\rho = M/N$. To quantify the diversity or uncertainty of the retained activations, we introduce the Shannon entropy S in terms of the activation density, as given in Eq. (2):

$$S = -\ln \rho. \tag{2}$$

As the threshold τ increases, the density ρ decreases, resulting in a corresponding reduction in entropy.

3.3 System Temperature

We define an effective temperature T in terms of the pruning threshold τ to quantify the intensity of pruning, as shown in Eq. (3):

$$T = \frac{1}{\tau}. \tag{3}$$

Here, T serves as a formal parameter that reflects system "hotness" (low τ, high T, dense activations) vs. "coldness" (high τ, low T, sparse activations). This analogy provides an intuitive interpretation of how thresholding governs information retention and controls network complexity.

3.4 Internal Energy

Let Σ denote the sum of absolute activation values after pruning, and let Σ_0 be the corresponding sum at $\tau = 0$. The internal energy E, defined in Eq. (4), captures the relative amount of information retained in the pruned network:

$$E = \ln\left(\frac{\Sigma}{\Sigma_0}\right). \tag{4}$$

This logarithmic ratio quantifies the remaining "energy" of the system and reflects how aggressively pruning suppresses the activation magnitude distribution.

3.5 Free Energy

The free energy F balances internal energy (E) and entropy (S), scaled by temperature (T), as given in Eq. (5):

$$F = E + T \cdot S. \tag{5}$$

Using the definitions above, we obtain the equivalent expression in Eq. (6):

$$F = E - T \cdot \ln\left(\frac{M}{N}\right). \tag{6}$$

Free energy serves as the central metric in our framework. Its extrema and curvature encode characteristic signatures of phase transitions—points at which pruning shifts from benign sparsification to rapid accuracy degradation. These dynamics and the associated critical thresholds are examined in detail in the following section.

3.6 Phase Transitions and Critical Threshold Identification via Free Energy Analysis

The free energy function $F(\tau)$ provides a principled mechanism for detecting phase transitions during activation pruning. In analogy to statistical physics—where phase transitions correspond to non-analytic behavior in thermodynamic potentials—we treat the pruning threshold τ as a control parameter and examine how $F(\tau)$ evolves as τ increases. Empirically, we observe that network accuracy remains stable over a substantial range of thresholds but declines sharply beyond a critical point τ^*. This transition consistently coincides with characteristic features of the free-energy curve, typically an extremum, an inflection point, or a pronounced change in curvature. In practice, we identify τ^* by locating such features in $F(\tau)$, often corroborated by a zero-crossing or peak in the first derivative $\frac{dF}{d\tau}$.

Because extrema, inflection points, and curvature changes all signal the onset of thermodynamic instability, we select the earliest such feature encountered as τ increases, yielding a consistent and conservative estimate of the critical threshold τ^*.

This thermodynamic instability reflects a phase transition from a dense, information-rich activation regime to a sparse, degraded one. The critical threshold τ^* marks the point at which accuracy begins to deteriorate and thus provides an approximate indication of the safe pruning range. Importantly, τ^* is not defined by a predetermined accuracy tolerance; instead, it is inferred directly from the onset of the free-energy transition, offering a model-agnostic and architecture-independent criterion for determining pruning limits.

The associated critical sparsity β^*—defined as $1-\rho$ evaluated at τ^*—indicates the maximum safe sparsity level before performance degradation becomes severe. Thus, free-energy analysis provides a robust and theoretically grounded method for approximating pruning thresholds, reducing the dependence on exhaustive empirical tuning. The observed correspondence between accuracy loss and characteristic changes in $F(\tau)$ generalizes across architectures and datasets, thus enhancing the interpretability of pruning dynamics through the framework of phase transition theory.

3.7 Renormalized Free Energy for Efficient Threshold Selection

Building on the free-energy analysis used to identify critical thresholds, we now introduce a renormalization procedure that enables fast, architecture-agnostic estimation of pruning thresholds. This method operates on a fixed activation array—collected from the trained model and a representative dataset prior to pruning—and simulates the pruning process by applying magnitude thresholds directly to this array, treating activations as an unstructured distribution. Thermodynamic quantities, including density, Shannon entropy, internal energy, and free energy, are then computed using the same definitions introduced earlier, but without requiring repeated model inference. This strategy can substantially reduce computational cost while preserving the characteristic shapes and derivative patterns of the free-energy curves obtained from full evaluation.

Although the activation array depends on the specific model and dataset, the renormalized analysis abstracts away architectural details and eliminates the need for retraining, providing a general and efficient approach for approximating critical pruning thresholds. As demonstrated in Section 5, the renormalized free-energy curves closely match those derived from the full evaluation, confirming both their accuracy and reliability. This method thus elevates pruning-threshold selection from heuristic tuning to a principled, physics-inspired procedure.

Procedure for Determining the Critical Pruning Threshold

Since the renormalized free energy $F(\tau)$ reliably reproduces the characteristic shape and stability regions observed under full pruning, we determine the critical pruning threshold τ^* directly from the

renormalized curve. For a given model, we specify a step size $\Delta\tau$ and evaluate the renormalized free energy over a grid of thresholds $\tau \in [\tau_{\min}, \tau_{\max}]$. The resulting free-energy profile $F(\tau)$ is then analyzed together with its first discrete derivative $\Delta F(\tau) = F(\tau + \Delta\tau) - F(\tau)$.

The critical threshold τ^* is identified as the earliest point at which the curve exhibits a qualitative transition, such as an extremum or a pronounced change in slope. This point corresponds to the onset of accuracy degradation observed during full pruning, and the associated sparsity level is reported as $\beta^* = 1 - \rho(\tau^*)$. Because this procedure relies on renormalized free energy, it avoids the repeated inference required in full-evaluation experiments and thus provides a substantially faster and architecture-agnostic method for locating the critical pruning threshold.

4 Experimental Setup

4.1 Datasets and Preprocessing

To evaluate our proposed method, we conducted experiments on both image and text classification tasks. For image classification, we used CIFAR-10 (50,000 training images, 10,000 test images, 10 classes) and the Oxford-102 Flowers dataset (8189 images, 102 classes; split into 1020 training, 1020 validation, and 6149 test samples). For text classification, we used AG News (120,000 training samples, 7600 test samples, 4 classes) and Yelp Review Full (650,000 training samples, 50,000 test samples, 5 sentiment categories).

CIFAR-10 images were normalized to the range $[-1, 1]$ using a channel-wise mean and standard deviation of 0.5, with a batch size of 64. Oxford-102 images were resized to 64×64 pixels and normalized using ImageNet statistics, with a batch size of 32.

Text datasets were tokenized, padded, and truncated to fixed maximum sequence lengths determined from training-data percentiles. Transformer-based models used their pretrained subword tokenizers, whereas the LSTM employed an NLTK-based word-level tokenizer. Text was converted to lowercase when appropriate, and special tokens (padding, unknown) were inserted as required. For GPT-2, the end-of-sequence token was used as padding. Tokenized inputs (IDs and attention masks) were processed using a batch size of 32.

4.2 Model Architectures

We evaluated a range of architectures spanning both non-pretrained and pretrained models. For image classification, the non-pretrained models include an MLP, implemented as a fully connected network with three hidden layers and pruning applied after ReLU activations, and a CNN consisting of three convolutional layers with batch normalization, ReLU, and max-pooling, followed by fully connected layers, with pruning applied after the convolutional layers and the first two fully connected layers. The pretrained models include ResNet-18 [24], with pruning applied after ReLU activations in the residual blocks; MobileNetV2 [25], which uses inverted residual blocks and pruning applied after ReLU6 activations; and a lightweight Vision Transformer (ViT) [26], with pruning applied after linear layers and GELU activations in the MLP blocks, while the last four transformer blocks and the classification head are fine-tuned.

For text classification, we evaluated one non-pretrained LSTM model and four pretrained transformer models covering encoder-only, encoder–decoder, and decoder-only architectures. The LSTM includes an embedding layer, a unidirectional LSTM, and a three-layer classifier, with pruning applied after the intermediate classifier layers. BERT [27] and ELECTRA [28] are encoder-only models, with pruning applied to the intermediate fully connected layers within the encoder blocks and to the classifier, using hooks on the encoder layers. T5 [29], an encoder–decoder model, is pruned after the fully connected layers in both encoder

and decoder blocks and in the classifier, using the pooled mean of the final decoder states. GPT-2 [30], a decoder-only model, is pruned after the fully connected layers in the decoder blocks and in the classifier.

4.3 Pruning Procedure and Parameter Settings

All models were trained using the Adam optimizer with a learning rate of 0.001 and cross-entropy loss. Dropout with a rate of 0.3 was applied to fully connected layers. Early stopping based on validation accuracy, with a patience of three epochs, was used across all models to mitigate overfitting. Training was conducted on a single GPU with CUDA support, and a fixed random seed was used to ensure reproducibility. No fine-tuning or additional optimization techniques were employed, as our objective was to analyze pruning behavior rather than to maximize accuracy.

After training, activation pruning was applied over a range of thresholds. For each threshold, test accuracy and activation statistics were recorded to compute the thermodynamic quantities (density, entropy, internal energy, and free energy) defined in Section 3. The implementation of our pruning framework, along with scripts for reproducing all experiments and free-energy analyses, is available at https://github.com/hse-scila/Activation_Pruning (accessed on 12 November 2025).

5 Results and Analysis

For each model–dataset pair, we present six plots that show test accuracy, free energy, and its derivative as functions of the pruning threshold (τ) and sparsity (β). Subfigures (a)–(c) display these metrics with respect to τ, while subfigures (d)–(f) present the same quantities as functions of β, the fraction of pruned activations. On the free-energy and derivative plots, we overlay the curves produced by the renormalized free-energy method (Sections 3.6 and 3.7) together with the full-evaluation results. Throughout the figures, blue curves correspond to full evaluation (empirical free energy), and pink curves correspond to the renormalized method (statistical free energy). This comparison shows that the renormalized curves preserve the characteristic phase-transition patterns, with critical thresholds closely matching those obtained by full evaluation. In some models, the curves nearly coincide; in others, they exhibit a small vertical offset, yet the estimated critical values remain well aligned.

When comparing free-energy curves across methods, we emphasize the shape of the curve rather than its absolute magnitude. Vertical shifts may arise from renormalization constants and do not influence the derivative-based transition detection. We therefore define a "match" as the agreement of the first qualitative transition (extremum or inflection point) at the same sparsity level β^*. Under this criterion, the renormalized and full-evaluation free-energy curves consistently match across all examined architectures.

Furthermore, for all results reported below, the critical thresholds and sparsity levels are stable across random seeds, with variability typically within 3%–5%. Representative mean values are shown for clarity and conciseness.

5.1 Vision Models

We begin by analyzing our framework on vision models and datasets. On CIFAR-10 (Fig. 2), the MLP maintains an accuracy of approximately 58% under moderate pruning thresholds ($\tau \approx 0.36$–0.40, $\beta \approx 0.32$–0.40), after which accuracy declines steadily. The full-evaluation free energy reaches its peak earlier ($\tau \approx 0.12$, $\beta \approx 0.15$), providing an early indication of impending instability well before the accuracy drop becomes pronounced. Although the free-energy extremum does not coincide exactly with the point where accuracy begins to fall, it provides a principled approximation of the critical region, as clearly visible from the plotted curves. This transition is further supported by the derivative curves ($dF/d\tau$ and $dF/d\beta$), and the sparsity

curves exhibit consistent patterns. On Flowers-102 (Fig. 3), a similar trend emerges: accuracy remains largely stable (around 31%) up to moderate thresholds, while peaks and curvature changes in the free-energy curves signal the onset of instability. The renormalized free-energy curves follow the same qualitative behavior, with only a slight vertical shift, and accurately capture the critical region for both datasets.

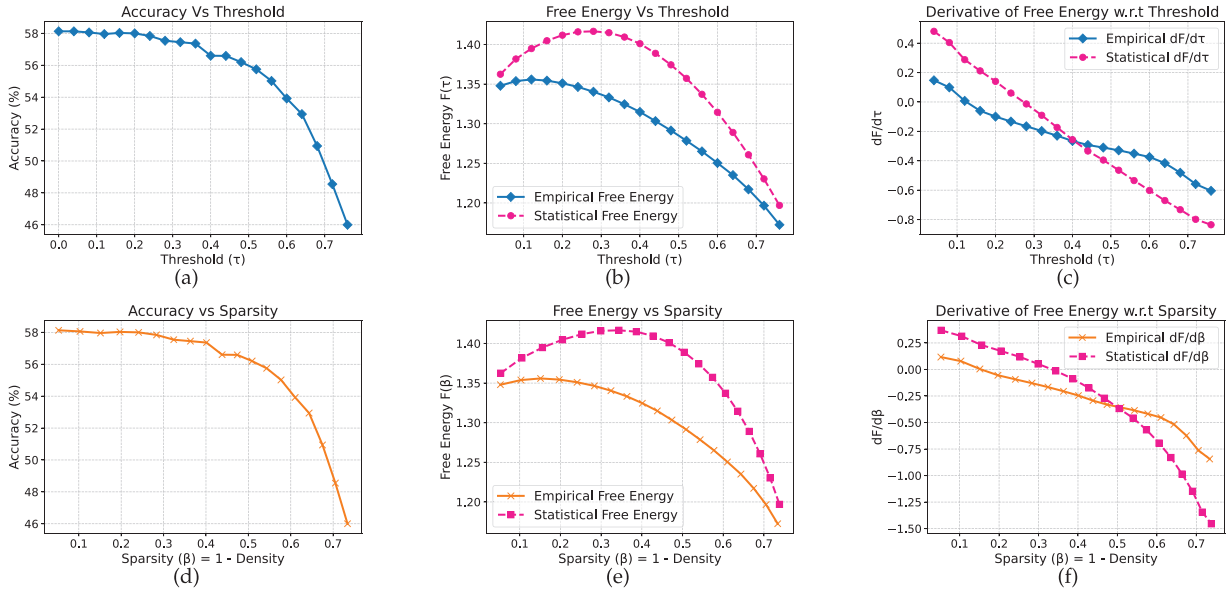


Figure 2: MLP model on CIFAR-10 dataset

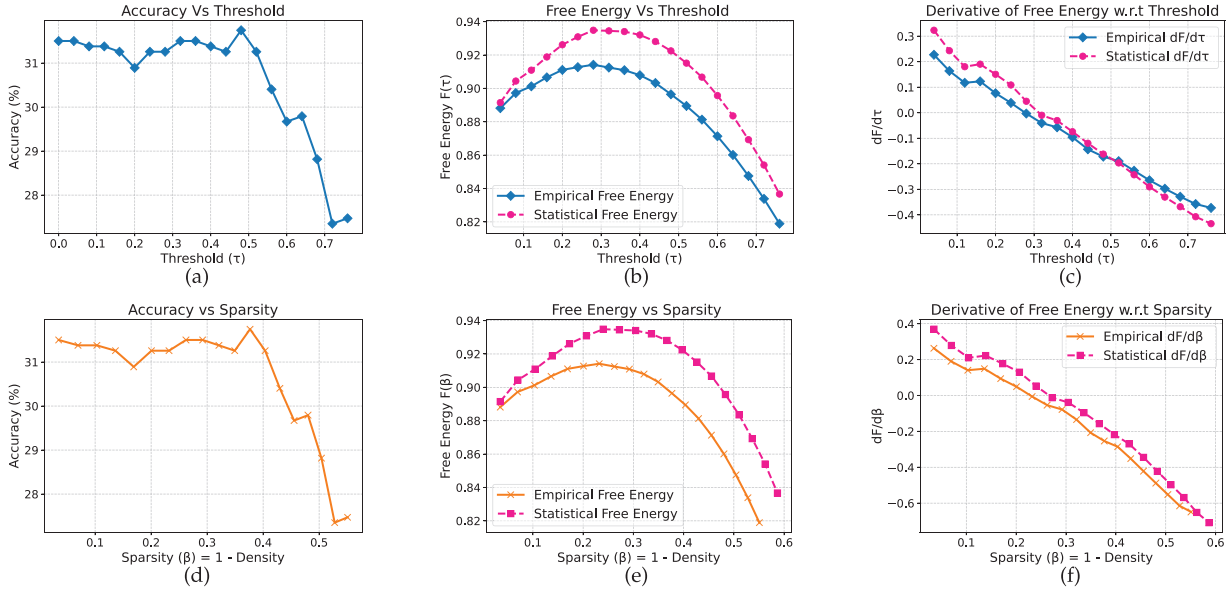


Figure 3: MLP model on Flowers-102 dataset

For the convolutional network, trends similar to those observed for the MLP arise across both datasets. On CIFAR-10 (Fig. 4), accuracy remains near 76% for moderate pruning thresholds ($\tau \approx 0.06$–0.08, $\beta \approx 0.23$–0.29) before gradually declining to about 70%. The full-evaluation free energy again reaches its peak slightly earlier, and the derivative curves highlight a clear phase transition within this region. On Flowers-102

(Fig. 5), accuracy remains stable at approximately 43%–44% up to thresholds of $\tau \approx 0.24$–0.30 ($\beta \approx 0.18$–0.23), after which it declines more sharply. The renormalized free-energy curves mirror the behavior of the full-evaluation curves, capturing the same critical trends with consistent qualitative alignment.

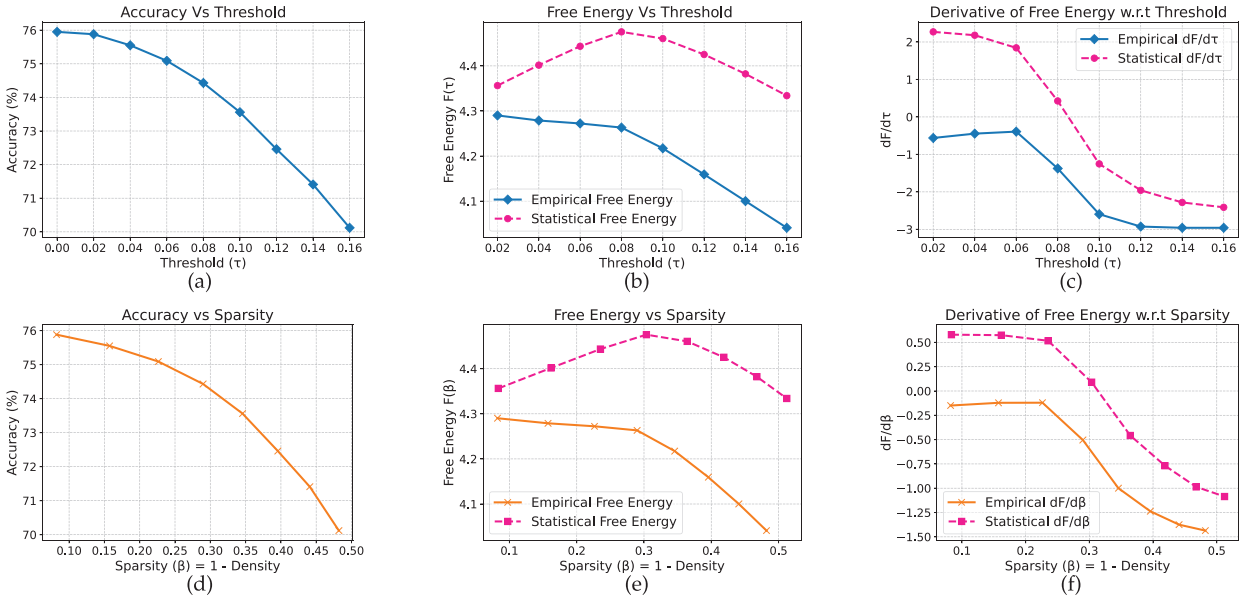


Figure 4: CNN model on CIFAR-10 dataset

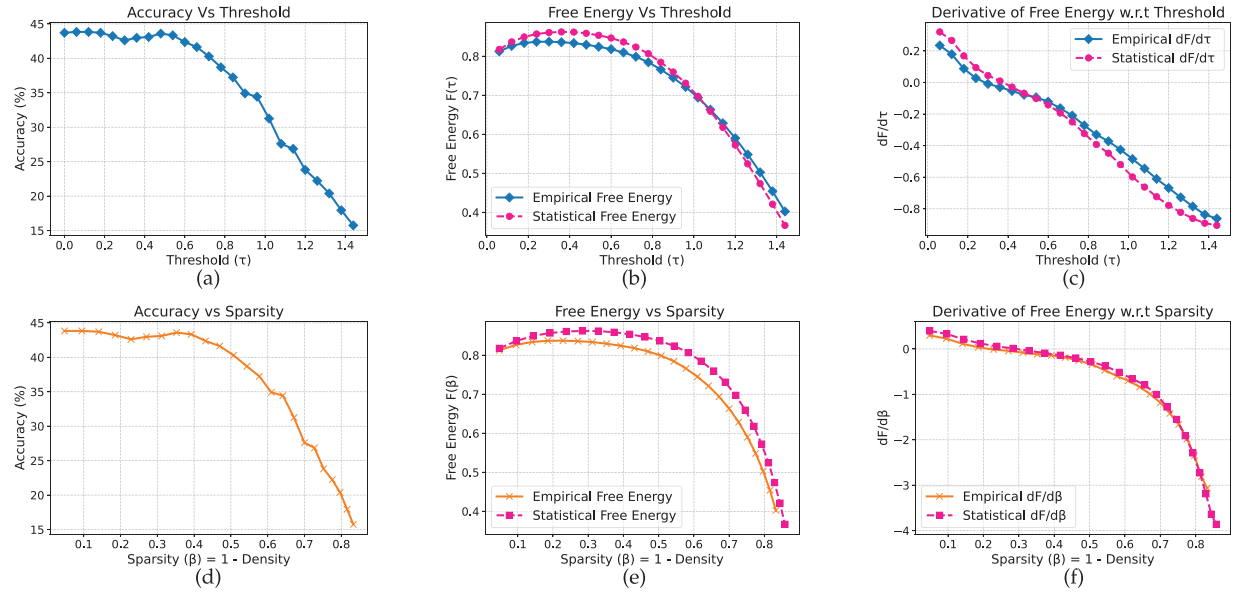


Figure 5: CNN model on Flower-102 dataset

Compared to shallow MLP and CNN, ResNet-18 exhibits a noticeably sharper transition. On CIFAR-10 (Fig. 6), accuracy remains close to 79% before declining markedly between $\tau \approx 0.20$–0.30. The full-evaluation free energy increases steadily and reaches a peak around $\tau \approx 0.30$–0.33 ($\beta \approx 0.50$–0.55), with the derivative curves clearly indicating the transition. On Flowers-102 (Fig. 7), accuracy remains stable at roughly 65%

initially and begins to decline between $\tau \approx 0.20$–0.30, while the free energy peaks near $\tau \approx 0.42$ ($\beta \approx 0.59$). The sharpness of the transition is also reflected in the sparsity curves. In contrast to the earlier MLP and CNN results, the free-energy peaks for ResNet-18 occur closer to the thresholds at which accuracy begins to deteriorate, providing a more precise indication of the critical region. The renormalized free-energy curves closely follow the full-evaluation trends, accurately capturing the transition behavior.

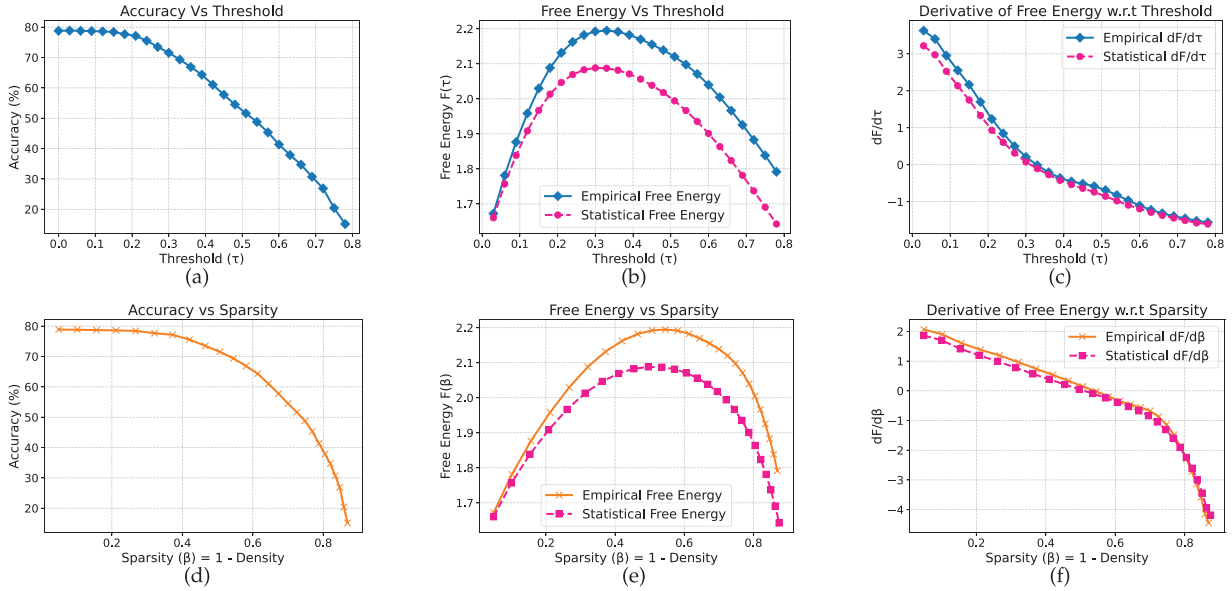


Figure 6: Resnet-18 model on CIFAR-10 dataset

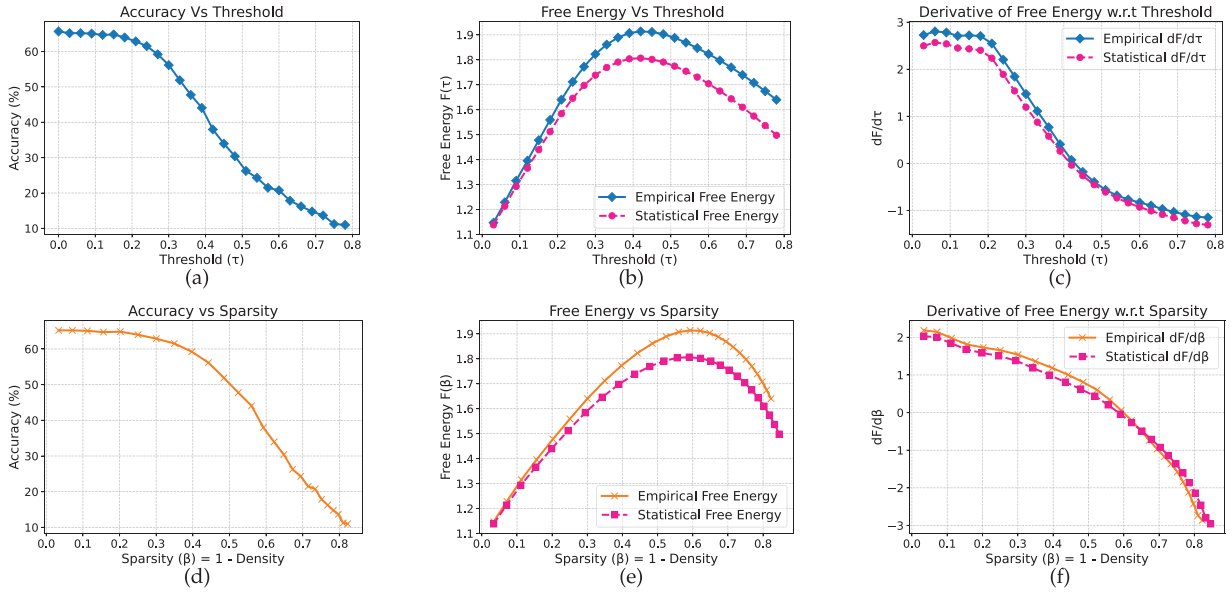


Figure 7: Resnet-18 model on Flower-102 dataset

For MobileNetV2, the results exhibit clearer dataset-dependent variation compared to the previous convolutional models. Moreover, unlike the earlier architectures that display a distinct peak in the free-energy curve, MobileNetV2 shows a bend on CIFAR-10 (Fig. 8) and a local minimum on Flowers-102 (Fig. 9).

On CIFAR-10, accuracy remains near 73% at low thresholds and begins to decline around $\tau \approx 0.12$ ($\beta \approx 0.21$), while the free energy shows a bend at $\tau \approx 0.06$ ($\beta \approx 0.12$), a feature confirmed by derivative analysis. On Flowers-102, accuracy stays around 74% up to $\tau \approx 0.10$ before decreasing sharply at $\tau \approx 0.22$ ($\beta \approx 0.35$). The free energy reaches a local minimum at $\tau \approx 0.12$ ($\beta \approx 0.19$), corroborated by a zero-crossing in the derivative. The renormalized free-energy curves reproduce these behaviors effectively, with only minor dataset-dependent deviations.

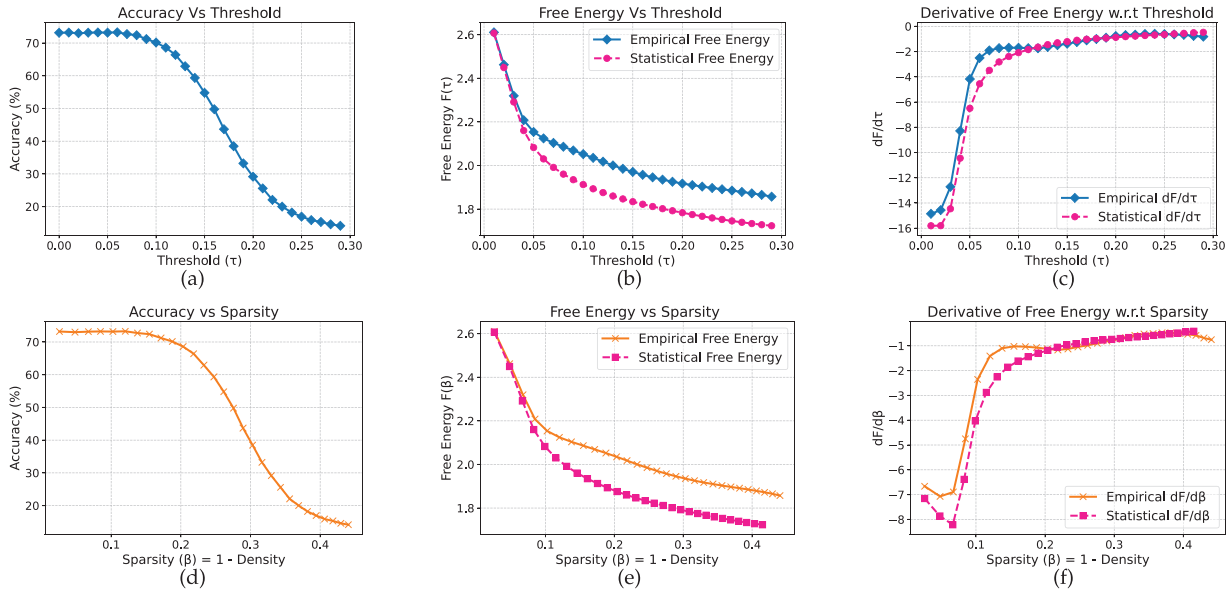


Figure 8: MobileNet model on CIFAR-10 dataset

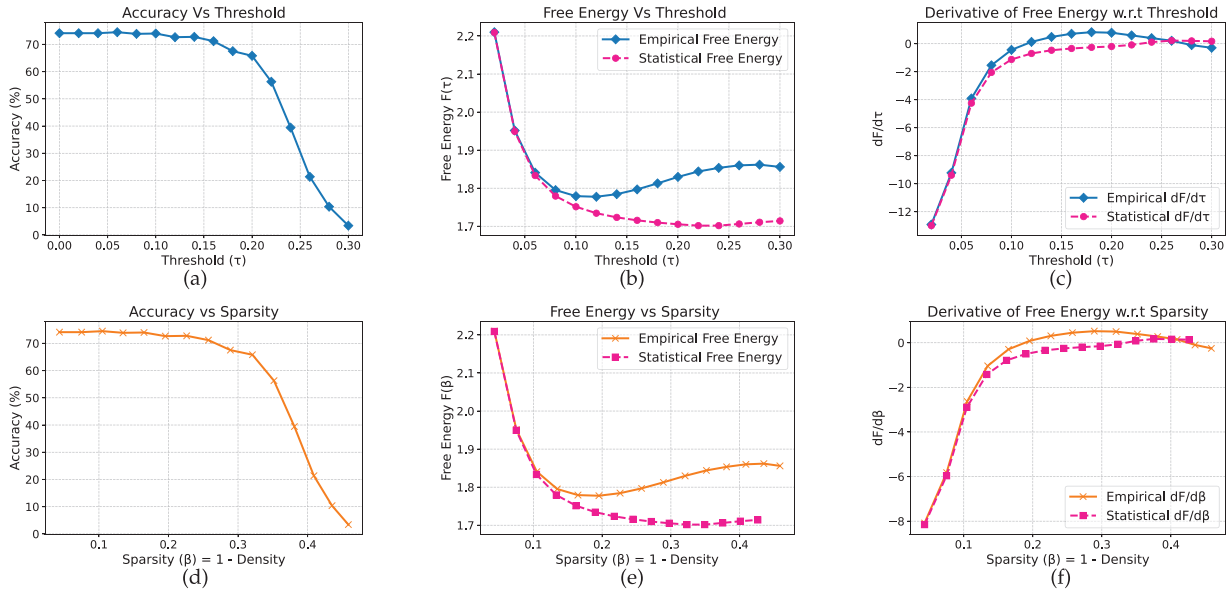


Figure 9: MobileNet model on Flower-102 dataset

Analysis of the Vision Transformer reveals qualitatively distinct behavior, characterized by a free-energy valley rather than a peak or bend. On CIFAR-10 (Fig. 10), accuracy remains near 97% at low thresholds and

begins to decline around $\tau \approx 0.13$ ($\beta \approx 0.66$), while the free-energy curve exhibits a clear valley at $\tau \approx 0.09$ ($\beta \approx 0.52$), accompanied by zero-crossings in the derivative. On Flowers-102 (Fig. 11), a similar valley appears at $\tau \approx 0.09$ ($\beta \approx 0.50$), preceding the onset of accuracy degradation. Notably, the renormalized free-energy curves show near-perfect overlap with the full-evaluation curves for both datasets, accurately tracking the critical region without significant shift.

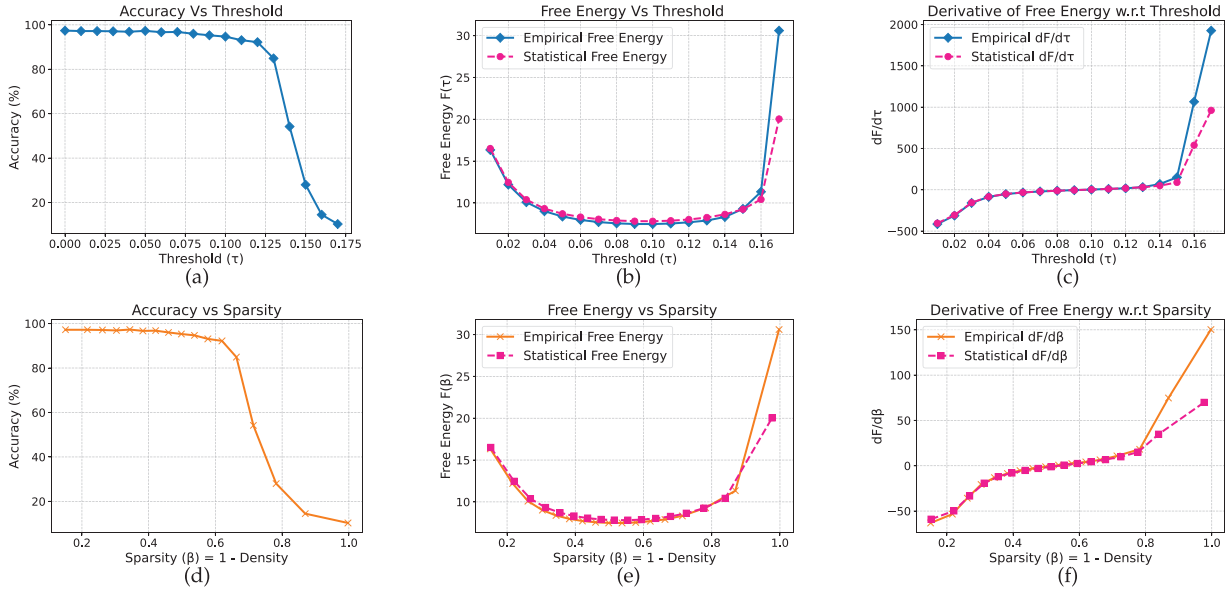


Figure 10: ViT model on CIFAR-10 dataset

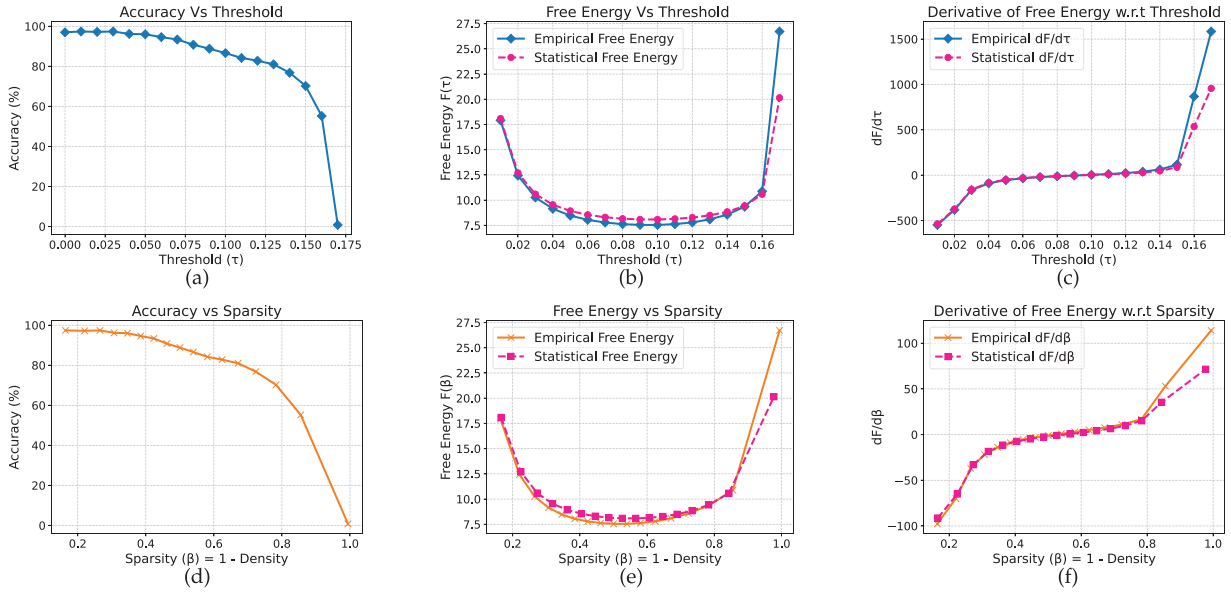


Figure 11: ViT model on Flower-102 dataset

Across the vision models, free energy tends to precede the onset of accuracy degradation for scratch-trained architectures (MLP, CNN) and roughly coincides with it for pretrained models (ResNet-18, MobileNetV2, ViT). Transitions in pretrained models appear smoother and more predictable, whereas

the MLP and CNN exhibit sharper accuracy drops and slightly noisier free-energy curves. Characteristic patterns—such as the bends in MobileNetV2 and the valley-shaped curve of the Vision Transformer—highlight the diversity of pruning dynamics across architectures. The renormalized free-energy curves generally track the full-evaluation trends, capturing the critical regions with only minor vertical offsets.

The differences in the shapes of the free-energy curves across architectures may be related to how information is distributed within each model. In CNNs and ResNets, activations within a channel tend to be spatially correlated, and a single channel often encodes a coherent local feature. Consequently, removing a channel during pruning can cause a sharp disruption in the learned representation, which manifests as a pronounced peak in the free-energy curve.

In MobileNetV2, the use of depthwise-separable convolutions and inverted residual blocks leads to more evenly distributed information and weaker inter-channel correlations. As a result, the loss of individual channels produces a smoother degradation of representation, reflected in a flatter free-energy curve.

In Vision Transformers, the self-attention mechanism enables global information flow across tokens, allowing the model to compensate for suppressed activations during the early stages of pruning. This flexibility gives rise to a broader free-energy valley that precedes the accuracy drop.

Table 1 summarizes the critical thresholds (τ^*) and sparsities (β^*) derived from accuracy degradation (ground truth) and from the full-evaluation free-energy analysis (thermodynamic prediction). Additional rows indicate whether the free-energy curve exhibits an extremum and whether a derivative sign change occurs at the critical point. The renormalized free energy is not included in the table, as its behavior is already clearly illustrated in the figures and discussed earlier, and its omission avoids unnecessary expansion of the table.

Table 1: Critical thresholds and sparsities from accuracy degradation and full evaluation free energy across vision models

Metric	MLP		CNN		ResNet		MobileNet		ViT	
	Acc.Based	FE.Based	Acc.Based	FE.Based	FAcc.Based	FE.Based	FAcc.Based	FE.Based	Acc.Based	FE.Based
CIFAR-10										
τ^*	≈ 0.36–0.40	≈ 0.12	0.06–0.08	0.02–0.04	0.20–0.30	0.30–0.33	0.12	0.06	0.13	0.09
β^*	≈ 0.32–0.40	≈ 0.15	0.23–0.29	0.04	0.50–0.55	0.05–0.10	0.21	0.12	≈ 0.66	≈ 0.52
Derivative = 0/Extr.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE Extr.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Flowers-102										
τ^*	0.40–0.44	0.28	0.24–0.30	0.18–0.24	0.20–0.30	0.42	0.22	0.12	≈ 0.13	≈ 0.09
β^*	0.32–0.35	0.23	0.18–0.23	0.14–0.19	0.35	0.59	0.35	0.19	0.69	0.56
Derivative = 0/Extr.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
FE Extr.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Time Efficiency Analysis

We evaluated the computational gains of the renormalized free energy (RFE) method relative to the full-evaluation free energy (FE) across all architectures. RFE provides substantial speedups: over 550× for the MLP on Flowers-102, between 4–19× for the CNN and MobileNet, and approximately 4× (CIFAR-10) to 8× (Flowers-102) for ResNet-18. The Vision Transformer exhibits mixed behavior—slightly slower on CIFAR-10 but nearly 2× faster on Flowers-102—reflecting dataset- and model-dependent computational characteristics.

Overall, these results show that RFE can significantly accelerate pruning analysis across a wide range of architectures while still reliably capturing the critical phase transition. Detailed timings and speedup factors are reported in Table 2.

Table 2: Computation time (seconds) and speedup factors for renormalized free energy compared to full evaluation across all vision models and datasets

	CIFAR-10					Flowers-102				
	MLP	**CNN**	**ResNet**	**MobileNet**	**ViT**	**MLP**	**CNN**	**ResNet**	**MobileNet**	**ViT**
Full Evaluation Time (s)	113.24	37.23	143.14	247.79	358.70	66.06	96.70	109.04	82.22	1122.09
Renormalized Time (s)	1.22	5.24	37.84	63.92	812.23	0.12	5.21	13.61	11.23	611.96
Speedup (×)	92.81	7.11	3.78	3.88	–	550.50	18.56	8.01	7.32	1.80

5.2 Text Models

Extending our analysis to text models, we observe that for non-pretrained LSTMs, the free-energy peaks align closely with the accuracy drop, in contrast to the scratch-trained vision models where the peak typically appears earlier. On AG News (Fig. 12), accuracy remains stable up to approximately $\tau \approx 0.42$ ($\beta \approx 0.32$), after which it declines sharply, defining the accuracy-based critical threshold. The full-evaluation free energy exhibits a peak near this point, with its derivative clearly indicating the transition, while the renormalized free-energy curve closely overlaps and slightly precedes the peak, providing an approximate estimate of $\tau^* \approx 0.42$ ($\beta^* \approx 0.32$).

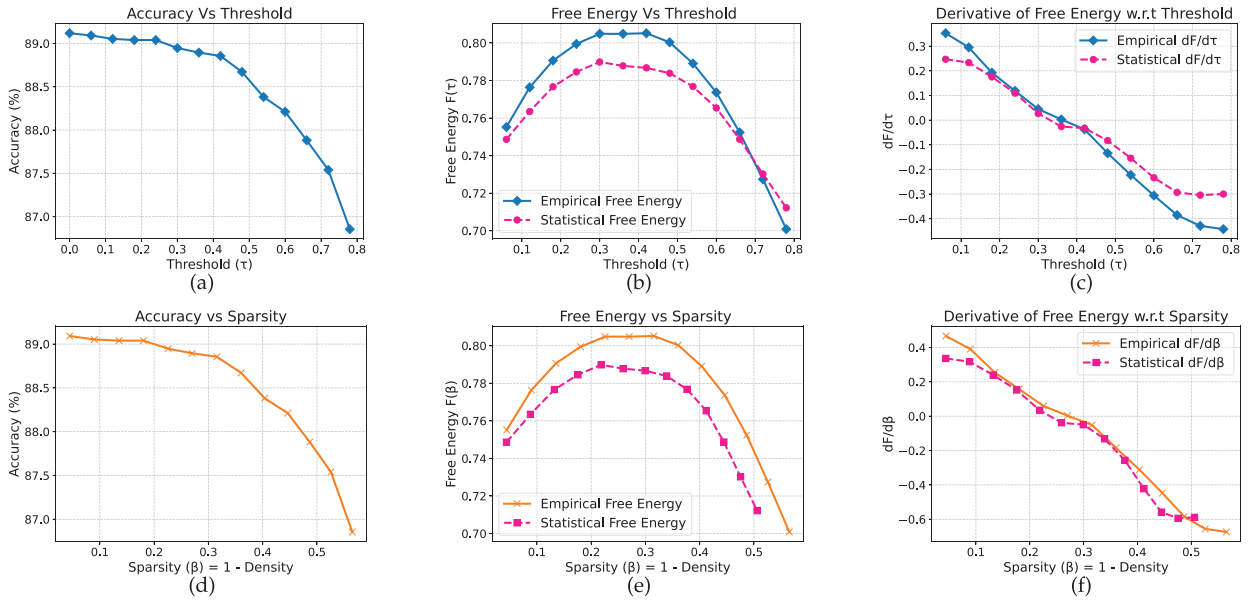


Figure 12: LSTM model on AG_News dataset

On Yelp (Fig. 13), a similarly abrupt transition is observed: accuracy remains near 62.9% before dropping between $\tau \approx 0.24$–0.30 ($\beta \approx 0.24$), with both the full and renormalized free-energy curves offering consistent guidance toward a critical threshold around $\tau^* \approx 0.30$. These findings indicate that non-pretrained text

models exhibit sharp pruning-induced transitions, reflecting limited robustness and a strong correspondence between free-energy indicators and accuracy degradation.

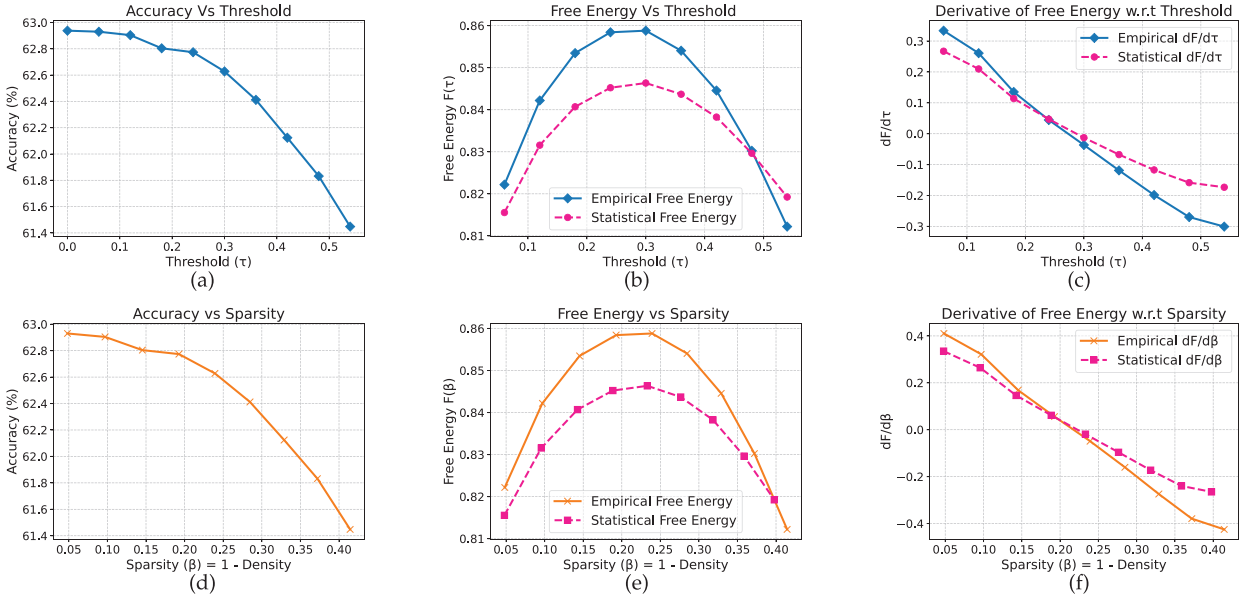


Figure 13: LSTM model on yelp full review dataset

In contrast, pretrained encoder-only transformers exhibit more structured and predictable behavior. On AG News (Fig. 14), BERT maintains approximately 93%–94% accuracy up to $\tau \approx 0.10$ ($\beta \approx 0.68$), during which the free energy decreases steadily toward a minimum. Between $\tau \approx 0.12$–0.18, accuracy collapses sharply as free energy rises, marking the transition. Interestingly, beyond $\tau \approx 0.20$ ($\beta \gtrsim 0.97$), accuracy begins to recover slightly (from 42% to 47%), accompanied by a second decline in free energy, suggesting a secondary stabilization regime. The derivative curves highlight the sharp changes around $\tau \approx 0.10$–0.12 as the onset of the critical region, in agreement with the renormalized free-energy evaluation.

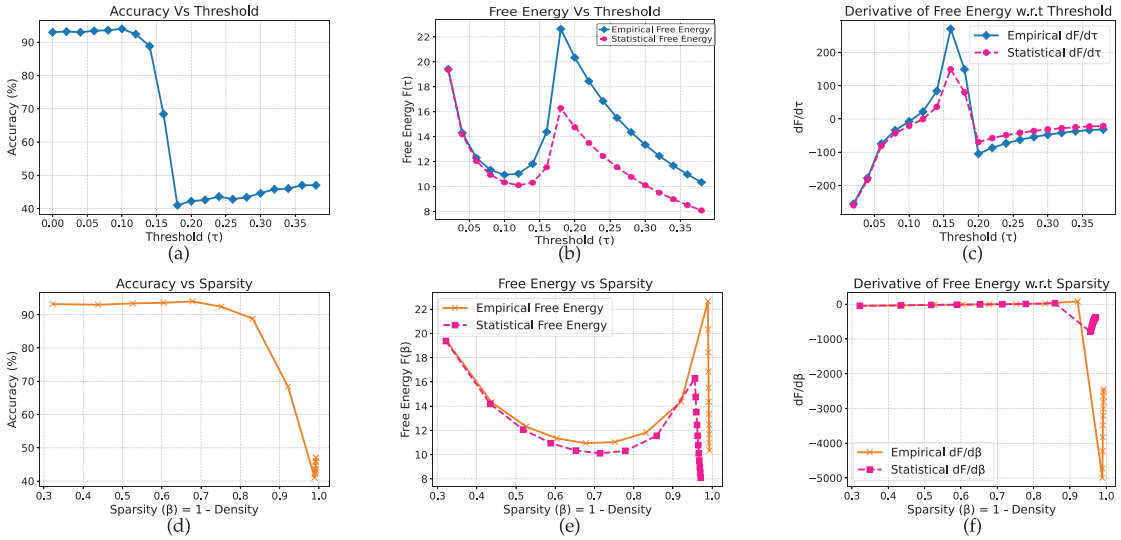


Figure 14: BERT model on AG_News dataset

On Yelp (Fig. 15), a similar accuracy drop occurs between $\tau \approx 0.12$–0.16, with free-energy minima approximately coinciding with the critical threshold. Both derivative and renormalized analyses support these points, underscoring the stabilizing effect of pretraining.

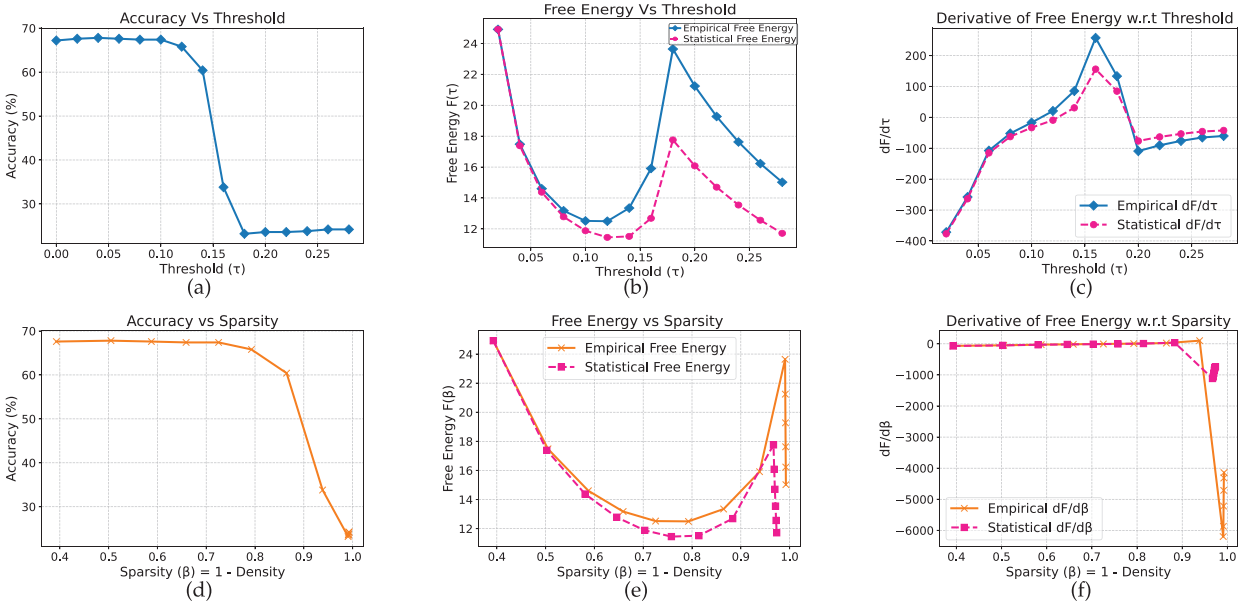


Figure 15: BERT model on yelp full review dataset

ELECTRA (Figs. 16 and 17) exhibits analogous behavior, with clear stable, transition, and collapsed regimes across both datasets. The full and renormalized free-energy curves yield consistent estimates of the critical threshold near $\tau^* \approx 0.10$ ($\beta^* \approx 0.55$), demonstrating reproducible phase-transition patterns in pretrained encoder-only models.

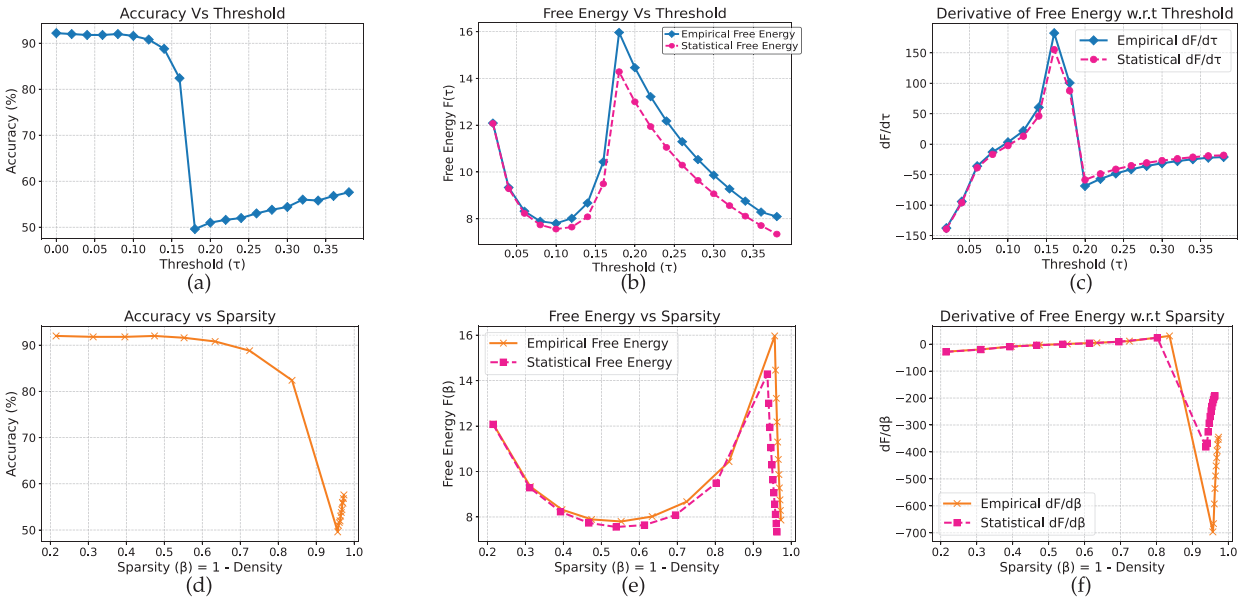


Figure 16: Electra model on AG_News dataset

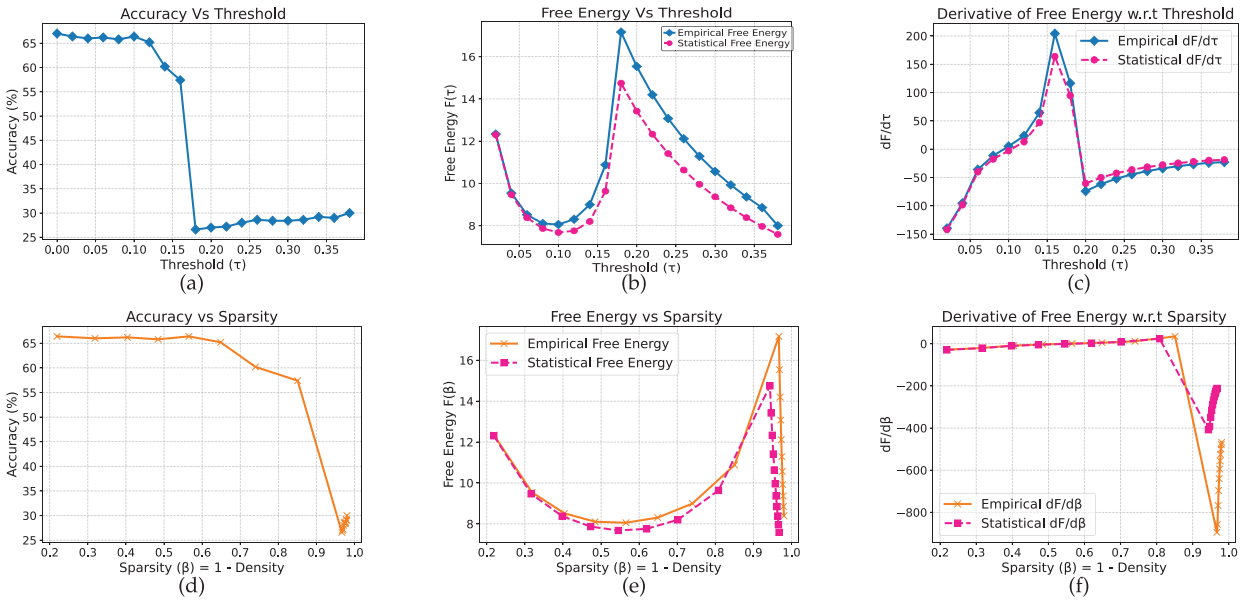


Figure 17: Electra model on yelp full review dataset

The results for T5, a sequence-to-sequence transformer, exhibit comparatively smoother pruning behavior. On AG News (Fig. 18), accuracy remains stable (approximately 91%–92%) up to $\tau \approx 0.8$ ($\beta \approx 0.045$), after which it begins to decline gradually. The corresponding free-energy curve displays a gentle maximum near these thresholds rather than the sharp extrema observed in other models. On Yelp (Fig. 19), a similar pattern appears, with a mild maximum around $\tau \approx 0.6$–0.8 ($\beta \approx 0.04$–0.06). These results highlight a model-dependent variation: sequence-to-sequence transformers tend to degrade more gradually under pruning, forming soft critical regions rather than abrupt phase-transition points.

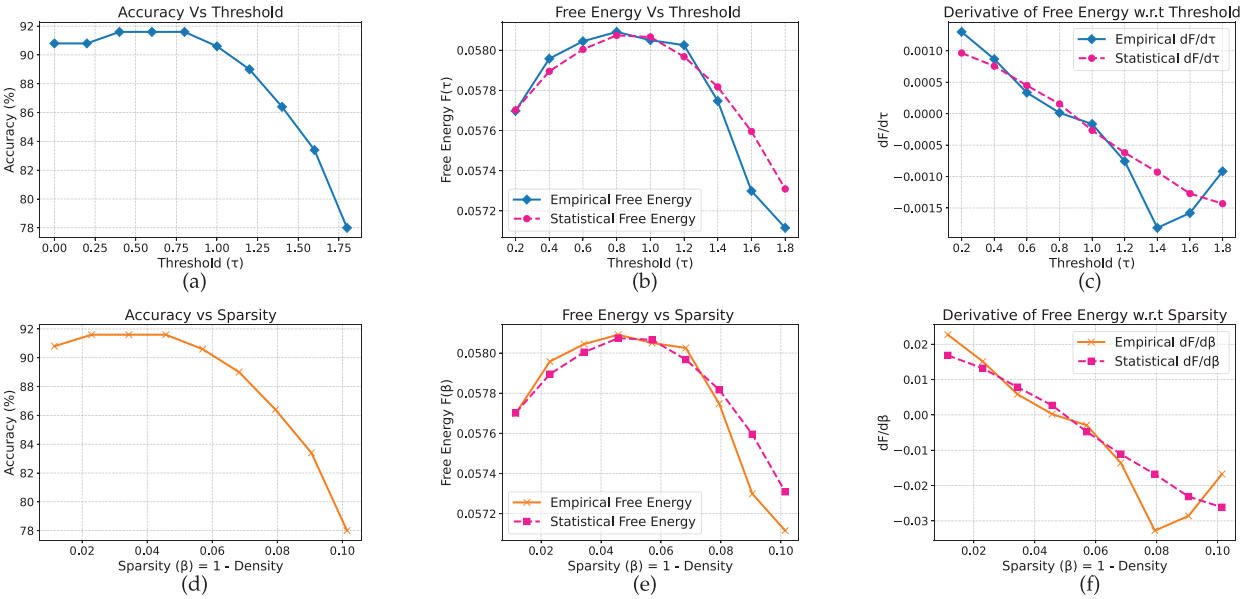


Figure 18: T5 model on AG_News dataset

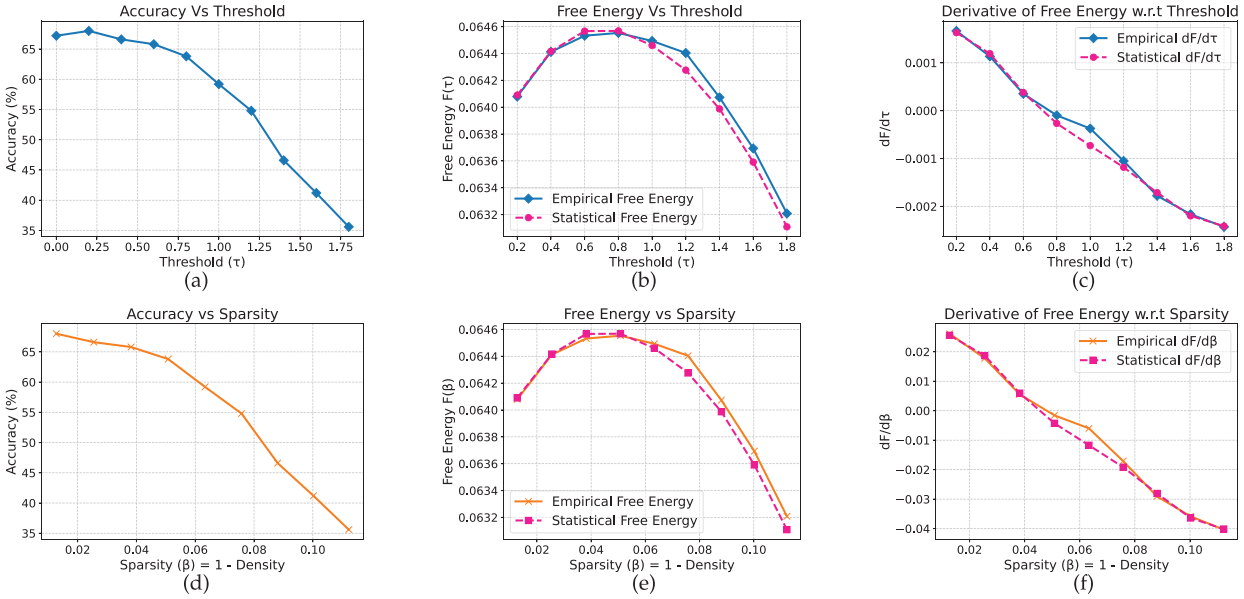


Figure 19: T5 model on yelp full review dataset

In GPT-2, the pruning dynamics vary noticeably across datasets. On AG News (Fig. 20), accuracy remains high (approximately 92%–94%) until a sharp decline at $\tau \approx 1.6$ ($\beta \approx 0.75$). The free-energy curve first exhibits a local minimum near $\tau \approx 1.0$ ($\beta \approx 0.46$), which does not correspond to the onset of accuracy degradation. Instead, the subsequent local maximum around $\tau \approx 1.6$ aligns closely with the abrupt accuracy drop, providing an approximate indication of the critical region beyond which the model can no longer maintain performance.

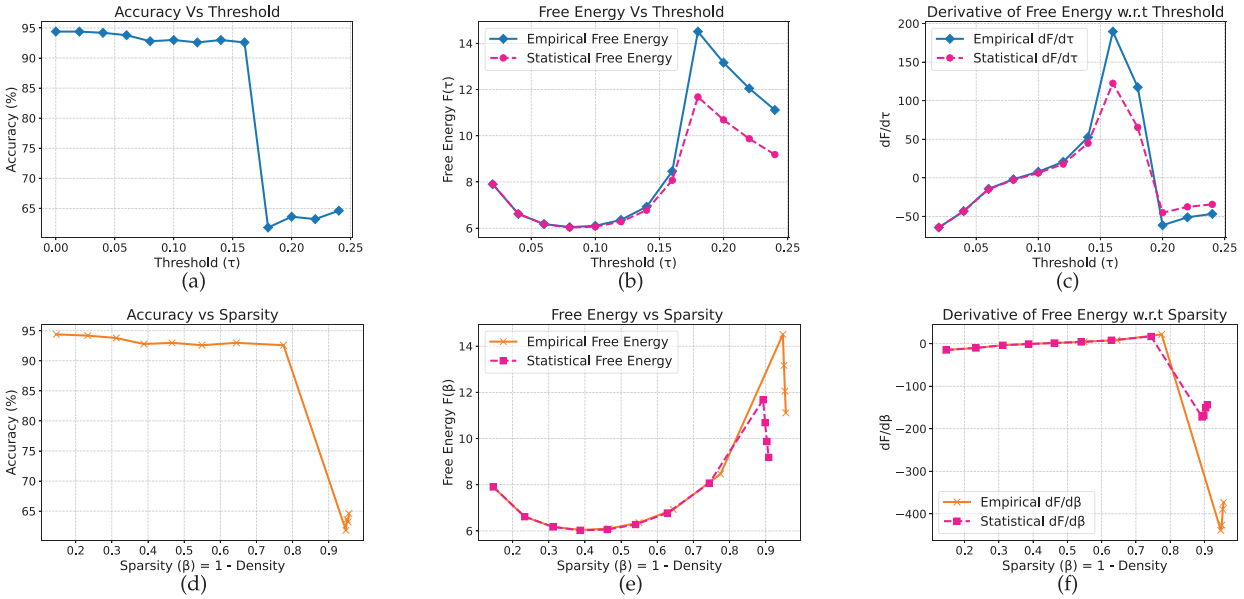


Figure 20: GPT_2 model on AG_News dataset

On Yelp (Fig. 21), accuracy declines earlier, at $\tau \approx 1.0$ ($\beta \approx 0.46$), with the free-energy minimum occurring near $\tau \approx 0.8$–1.0. The renormalized free-energy curves confirm these patterns, showing that

although the overall free-energy shapes are nearly identical across datasets, the relevant extremum depends on the dataset. This contrasts with the more consistent behavior observed in encoder-based models, where the same free-energy feature typically marks the transition.

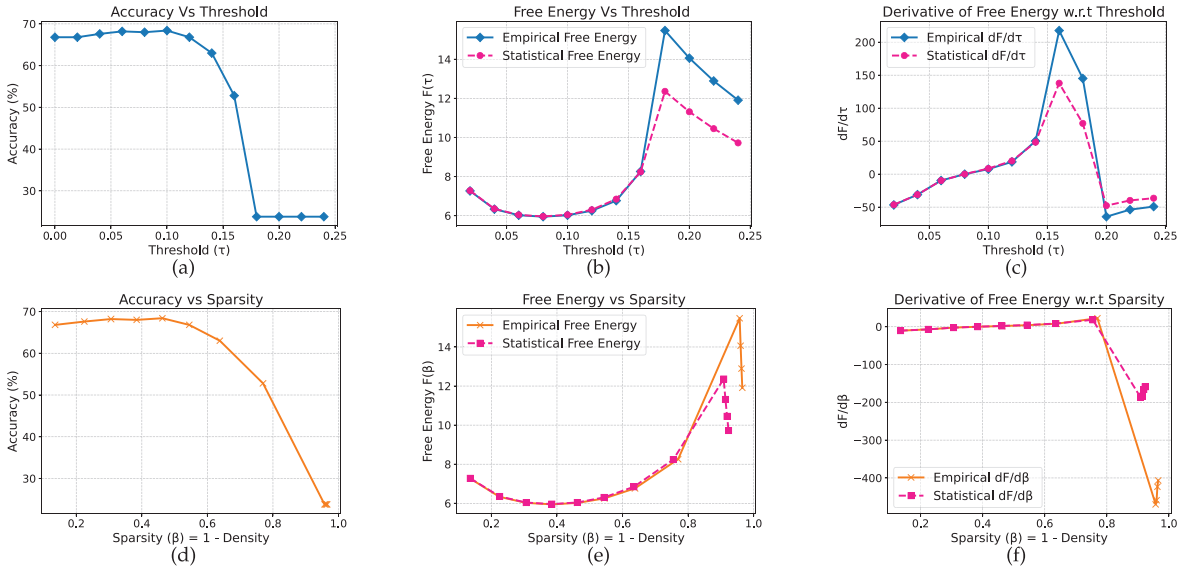


Figure 21: GPT_2 model on yelp full review dataset

Across the text models, free energy offers a reliable estimate of critical thresholds near the onset of accuracy degradation, with derivative curves marking the corresponding phase transitions. Non-pretrained LSTMs exhibit abrupt and largely dataset-independent transitions, whereas pretrained transformers display smoother and more predictable behavior. Encoder-only models (BERT, ELECTRA) show sharp minima that provide clear approximations of the critical threshold; T5 degrades gradually, forming soft critical regions; and GPT-2 demonstrates dataset-dependent extrema. In all cases, the renormalized free-energy curves closely follow the full-evaluation trends across architectures and datasets.

Table 3 summarizes the critical thresholds (τ^*) and sparsities (β^*) for all text models and datasets, extending the applicability of our framework beyond vision tasks and demonstrating its generality across modalities and model families.

Table 3: Critical thresholds and sparsities from accuracy degradation and full evaluation free energy across text models

Metric	LSTM		BERT		ELECTRA		T5		GPT-2	
	Acc.Based	FE Based	Acc.Based	FE Based	Acc.Based	FE Based	Acc.Based	FE Based	Acc.Based	FE Based
AG news										
τ^*	0.42	0.42	0.10–0.12	0.10–0.12	0.10	0.10	0.80	0.80	1.60	1.60
β^*	0.32	0.32	0.68–0.71	0.68–0.71	0.55	0.55	0.45	0.45	0.75	0.75
Derivative = 0/Peak	Yes	Yes	Yes	Yes	Yes	Yes	Soft	Soft	Yes	Yes
Free Energy Peak	Yes	Yes	Yes	Yes	Yes	Yes	Soft	Soft	Yes	Yes
Yelp review full										
τ^*	0.24–0.30	0.30	0.12	0.12	0.10	0.10	0.60	0.60–0.80	1.0	0.80–1.0

(Continued)

Table 3 (continued)

Metric	LSTM		BERT		ELECTRA		T5		GPT-2	
	Acc.Based	FE Based	Acc.Based	FE Based	Acc.Based	FE Based	Acc.Based	FE Based	Acc.Based	FE Based
β^*	0.24	0.23–0.24	0.76	0.76	0.55	0.55	0.04	0.04–0.06	0.46	0.38–0.46
Derivative = 0/Extr.	Yes	Yes	Yes	Yes	Yes	Yes	Soft	Soft	Yes	Yes
FE Extr.	Yes	Yes	Yes	Yes	Yes	Yes	Soft	Soft	Yes	Yes

Empirically, we observe that accuracy remains stable before the free-energy transition, typically varying by only 3%–5%. Once this transition is crossed, however, performance drops sharply (by at least 30% in our experiments). This behavior confirms that the free-energy transition corresponds to a phase-like boundary rather than a gradual degradation.

Time Efficiency Analysis

For text models as well, RFE provides substantial computational gains for simpler architectures. The LSTM achieves dramatic speedups—over 460× on AG News and 1700× on Yelp Review Full. Moderate improvements of 1.1–2.5× are observed for ELECTRA and T5, indicating consistent runtime reductions even for larger architectures. In contrast, BERT is slower than full evaluation on both datasets, while GPT-2 shows negligible speedup on Yelp and none on AG News. These results suggest that RFE can greatly accelerate pruning analysis for smaller or moderately sized text models, whereas its benefits are limited or absent for some larger pretrained transformers due to activation-processing overhead. Detailed timings and speedup factors are reported in Table 4.

Table 4: Computation time (seconds) and speedup factors for renormalized free energy compared to full evaluation across all text models and datasets

	AG news					Yelp review full				
	LSTM	BERT	T5	ELECTRA	GPT-2	LSTM	BERT	T5	ELECTRA	GPT-2
Full-Evaluation Time (s)	28.07	64.05	31.93	53.07	42.14	503.00	173.73	142.90	229.98	723.13
Renormalized Time (s)	0.06	98.61	16.04	47.87	110.01	0.29	448.92	56.92	189.96	649.55
Speedup (×)	468	—	1.99	1.11	—	1734	—	2.51	1.21	1.11

6 Conclusion

We presented a thermodynamics-inspired framework for activation pruning, using free-energy dynamics as a proxy for approximating critical pruning thresholds. Across both vision and language models, extrema and inflection points in the free-energy curve typically precede or closely coincide with the onset of accuracy degradation, revealing architecture- and dataset-dependent behavior: non-pretrained networks (MLP, CNN, LSTM) exhibit sharp, abrupt transitions, pretrained models such as ResNet-18, MobileNetV2, and ViT display smoother and more gradual dynamics, and transformer architectures show model- and dataset-specific extrema.

To improve efficiency, we introduced a renormalized free-energy method that reconstructs the same diagnostic curves using only unpruned activations, thereby eliminating the need for repeated inference.

This renormalized evaluation yields substantial speedups for fully connected, convolutional, and residual architectures, achieving up to a 500× reduction in computational cost. For transformer-based text models (e.g., BERT and GPT-2) and ViT, the speedup is more moderate—and in some cases negligible—due to the higher cost of activation collection and the reduced redundancy in channel structure. Nonetheless, the free-energy transition consistently provides a reliable indicator of the critical sparsity threshold across all model families.

This work bridges theoretical insight and practical application, demonstrating that free energy offers a principled and generalizable perspective on pruning behavior. While our experiments focus on classification tasks, extending this framework to generative models, reinforcement learning settings, and finer-grained layerwise analyses represents a promising direction for future research.

Acknowledgement: This research was supported in part through computational resources of HPC facilities at HSE University.

Funding Statement: This work is an output of a research project implemented as part of the Basic Research Program at HSE University

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Rayeesa Mehmood, Sergei Koltcov, Anton Surkov and Vera Ignatenko; methodology, Rayeesa Mehmood, Sergei Koltcov, Anton Surkov and Vera Ignatenko; software, Rayeesa Mehmood; validation, Rayeesa Mehmood, Sergei Koltcov, Anton Surkov and Vera Ignatenko; formal analysis, Rayeesa Mehmood and Sergei Koltcov; investigation, Rayeesa Mehmood, Vera Ignatenko and Anton Surkov; resources, Rayeesa Mehmood and Anton Surkov; data curation, Rayeesa Mehmood; writing—original draft preparation, Rayeesa Mehmood; writing—review and editing, Vera Ignatenko, Sergei Koltcov and Anton Surkov; visualization, Rayeesa Mehmood and Vera Ignatenko; supervision, Sergei Koltcov; project administration, Vera Ignatenko; funding acquisition, Sergei Koltcov. All authors reviewed the results and approved the final version of the manuscript

Availability of Data and Materials: The data that support the findings of this study are openly available in Github at https://github.com/hse-scila/Activation_Pruning (accessed on 12 November 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Abbreviations

MLP	Multi-Layer Perceptron
DNN	Deep Neural Network
FC	Fully Connected
CNN	Convolutional Neural Network
ResNet	Residual Network
MobileNet	Mobile-Optimized Convolutional Neural Network
ViT	Vision Transformer
LSTM	Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
T5	Text-to-Text Transfer Transformer
GPT-2	Generative Pretrained Transformer 2

References

1. Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient convnets. In: Proceedings of the 5th International Conference on Learning Representations (ICLR'17); 2017 Apr 24–26; Toulon, France. p. 1683–95.
2. Sun M, Liu Z, Bair A, Kolter JZ. 1A Simple and effective pruning approach for large language models. In: The Twelfth International Conference on Learning Representations; 2024 May 7–11; Vienna, Austria. p. 4640–62.
3. Han S, Pool J, Tran J, Dally W. Learning both weights and connections for efficient neural network. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in neural information processing systems. Vol. 28. Red Hook, NY, USA: Curran Associates, Inc.; 2015. p. 1135–43.
4. Neo M. A comprehensive guide to neural network model pruning [Internet]. 2024 [cited 2025 May 15]. Available from: https://www.datature.io/blog/a-comprehensive-guide-to-neural-network-model-pruning.
5. Heyman A, Zylberberg J. Fine granularity is critical for intelligent neural network pruning. Neural Comput. 2024;36(12):2677–709. doi:10.1162/neco_a_01717.
6. Chen E, Wang H, Shi Z, Zhang W. Channel pruning for convolutional neural networks using l0-norm constraints. Neurocomputing. 2025;636:129925. doi:10.1016/j.neucom.2025.129925.
7. Hu H, Peng R, Tai YW, Tang CK. Network trimming: a data-driven neuron pruning approach towards efficient deep architectures. arXiv:1607.03250. 2016.
8. Foldy-Porto T, Venkatesha Y, Panda P. Activation density driven energy-efficient pruning in training. arXiv:2002.02949. 2020.
9. xiao Li Z, You C, Bhojanapalli S, Li D, Rawat AS, Reddi SJ, et al. The lazy neuron phenomenon: on emergence of activation sparsity in transformers. arXiv:2210.06313. 2022.
10. Tang C, Sun W, Yuan Z, Liu Y. Adaptive pixel-wise structured sparse network for efficient CNNs. arXiv:2010.11083. 2021.
11. Hussien M, Afifi M, Nguyen KK, Cheriet M. Small contributions, small networks: efficient neural network pruning based on relative importance. arXiv:2410.16151. 2024.
12. Kurtz M, Kopinsky J, Gelashvili R, Matveev A, Carr J, Goin M, et al. Inducing and exploiting activation sparsity for fast inference on deep neural networks. In: Daumé H III, Singh A, editors. Proceedings of the 37th International Conference on Machine Learning. Vol. 119. London, UK: PMLR; 2020. p. 5533–43.
13. Ganguli T, Chong EK. Activation-based pruning of neural networks. Algorithms. 2024;17(1):48. doi:10.3390/a17010048.
14. Zhao K, Jain A, Zhao M. Adaptive activation-based structured pruning. arXiv:2201.10520. 2022.
15. Grimaldi M, Ganji DC, Lazarevich I, Deeplite SS. Accelerating deep neural networks via semi-structured activation sparsity. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway, NJ, USA: IEEE; 2023. p. 1171–80.
16. Wang M, Zhang M, Liu X, Nie L. Weight-aware activation sparsity with constrained bayesian optimization scheduling for large language models. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V, editors. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL; 2025. p. 1086–98.
17. Zhang Z, Liu Z, Tian Y, Khaitan H, Wang Z, Li S. R-Sparse: rank-aware activation sparsity for efficient LLM inference. arXiv:2504.19449. 2025.
18. Liu J, Ponnusamy P, Cai T, Guo H, Kim Y, Athiwaratkun B. Training-free activation sparsity in large language models. arXiv:2408.14690. 2025.
19. An T, Cai R, Zhang Y, Liu Y, Chen H, Xie P, et al. Amber pruner: leveraging n:m activation sparsity for efficient prefill in large language models. arXiv:2508.02128. 2025.
20. Ding Y, Chen DR. Optimization based layer-wise pruning threshold method for accelerating convolutional neural networks. Mathematics. 2023;11(15):3311. doi:10.3390/math11153311.
21. Sanh V, Wolf T, Rush AM. Movement pruning: adaptive sparsity by fine-tuning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 20378–89.

22. Frantar E, Alistarh D. SparseGPT: massive language models can be accurately pruned in one-shot. arXiv:2301.00774. 2023.
23. Frantar E, Ashkboos S, Hoefler T, Alistarh D. GPTQ: accurate post-training quantization for generative pre-trained transformers. arXiv:2210.17323. 2023.
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2016. p. 770–8.
25. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2018. p. 4510–20.
26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations. Red Hook, NY, USA: Curran Associates, Inc.; 2021. p. 1–21.
27. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL; 2019. p. 4171–86.
28. Clark K, Luong MT, Le QV, Manning CD. ELECTRA: pre-training text encoders as discriminators rather than generators. arXiv:2003.10555. 2020.
29. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
30. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners [Internet]. San Francisco, CA, USA: OpenAI; 2019 [cited 2025 Nov 7]. Available from: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.