POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Structure and Redundancy in Large Language Models: A Spectral Study via Random Matrix Theory

Tesi di Laurea Magistrale in
Computer Science Engineering - Ingegneria Informatica

Author: **Davide Ettori**

Student ID: 244696
Advisor: Prof. Marco Brambilla
Co-advisors: Prof. Amit Ranjan Trivedi
Academic Year: 2025-26

Abstract

This thesis addresses two persistent and closely related challenges in modern deep learning, reliability and efficiency, through a unified framework grounded in Spectral Geometry and Random Matrix Theory (RMT). As deep networks and large language models continue to scale, their internal behavior becomes increasingly opaque, leading to hallucinations, fragile generalization under distribution shift, and growing computational and energy demands. By analyzing the eigenvalue dynamics of hidden activations across layers and inputs, this work shows that spectral statistics provide a compact, stable, and interpretable lens on model behavior, capable of separating structured, causal representations from noise-dominated variability. Within this framework, the first contribution, **EigenTrack**, introduces a real-time method for detecting hallucinations and out-of-distribution behavior in large language and vision-language models. EigenTrack transforms streaming activations into spectral descriptors such as entropy, variance, and deviations from the Marchenko–Pastur baseline, and models their temporal evolution using lightweight recurrent classifiers, enabling early detection of reliability failures before they appear in model outputs while offering interpretable insight into representation dynamics. The second contribution, **RMT-KD**, presents a principled approach to compressing deep networks via random matrix theoretic knowledge distillation. By interpreting outlier eigenvalues in activation spectra as carriers of task-relevant information, RMT-KD progressively projects networks onto lower-dimensional subspaces through iterative self-distillation, yielding significantly more compact and energy-efficient models while preserving accuracy and dense, hardware-friendly structure. Together, these contributions establish spectral geometry as a coherent foundation for diagnosing uncertainty and guiding compression in large-scale neural networks, demonstrating how Random Matrix Theory provides mathematically grounded tools for building deep learning systems that are more reliable, efficient, and trustworthy in high-stakes settings.

Keywords: Large Language Models, Random Matrix Theory, Model Compression, Hallucination Detection

Abstract in lingua italiana

Questa tesi affronta due problemi chiave del deep learning contemporaneo, affidabilità ed efficienza, proponendo un quadro unificato basato sulla Geometria Spettrale e sulla Random Matrix Theory (RMT). Con l'aumento di scala complessità dei Large Language Models, il loro funzionamento interno diventa sempre meno trasparente, dando origine a fenomeni come allucinazioni, scarsa robustezza ai cambi di distribuzione e costi computazionali ed energetici sempre più elevati. Studiando le dinamiche spettrali delle attivazioni interne, questo lavoro mostra che le statistiche sugli autovalori offrono una descrizione compatta, stabile e interpretabile del comportamento dei modelli, capace di distinguere rappresentazioni informative e strutturate da componenti prevalentemente di rumuore. In questo contesto, il primo contributo, **EigenTrack**, introduce un metodo in tempo reale per il rilevamento di allucinazioni e comportamenti out-of-distribution in large language models e vision-language models. EigenTrack trasforma le attivazioni dei modelli in descrittori spettrali sugli autovalori, come entropia, varianza e deviazioni dalla legge di Marchenko–Pastur, e ne analizza l'evoluzione temporale tramite modelli ricorrenti, individuando precocemente situazioni di inaffidabilità prima che emergano negli output del modello. Il secondo contributo, **RMT-KD**, propone un approccio rigoroso alla compressione delle reti neurali basato sulla knowledge distillation, guidato dalla RMT. Interpretando gli autovalori outlier degli spettri di attivazione come portatori di informazione rilevante per l'obiettivo, RMT-KD riduce progressivamente la dimensionalità dei modelli tramite auto-distillazione iterativa, ottenendo modelli molto più compatti ed efficienti dal punto di vista energetico senza sacrificare l'accuratezza né la struttura densa, favorevole all'hardware. Questi contributi mostrano come la prospettiva spettrale possa costituire una base coerente sia per analizzare l'affidabilità sia per guidare la compressione dei modelli su larga scala, dimostrando che la RMT può tradursi in strumenti matematicamente solidi per costruire sistemi di deep learning più affidabili ed efficienti.

Parole chiave: Grandi Modelli Linguistici, Teoria delle Matrici Casuali, Compressione dei Modelli, Rilevamento delle Allucinazioni

Contents

Introduction

In recent years, deep learning and large language models have become central to advances in artificial intelligence. Transformer-based architectures now underpin applications ranging from natural language understanding and text generation to computer vision and multimodal reasoning. These models have shown remarkable capabilities, often surpassing human performance on benchmark tasks, and are increasingly being adopted in domains such as healthcare, law, finance, and education. Their rapid adoption highlights both their transformative potential and the growing reliance of modern society on machine intelligence.

However, the success of these systems comes with critical challenges. On the one hand, large language models often suffer from reliability issues: they can produce hallucinations, generate plausible but incorrect outputs, and fail dramatically when faced with out-of-distribution inputs. Such behavior undermines trust in AI, particularly in safety-critical contexts where decisions must be both accurate and justifiable. On the other hand, the efficiency of these models is a major obstacle. State-of-the-art architectures contain billions of parameters, requiring enormous computational resources for training and inference. This makes them difficult to deploy on edge devices, costly to maintain in large-scale production, and unsustainable in terms of energy consumption.

Addressing these dual challenges of reliability and efficiency requires new methods that are not only empirically effective but also grounded in mathematical principles. One promising perspective arises from spectral geometry and Random Matrix Theory (RMT), which provide a rigorous way to analyze the eigenvalue spectra of hidden activations in neural networks. Eigenvalue statistics can reveal whether a model is encoding structured, causal information or drifting toward noise-like behavior, offering a compact and interpretable description of its internal dynamics. By leveraging these spectral insights, it becomes possible to both diagnose failures in real time and guide principled approaches to model compression.

This thesis develops such a spectral framework for modern deep learning and large language models. It introduces two contributions built on RMT: EigenTrack, a method

for real-time hallucination and out-of-distribution detection, and RMT-KD, a knowledge distillation approach for efficient network compression. Together, these methods demonstrate that spectral analysis of neural representations can provide a unifying foundation for making AI systems simultaneously more trustworthy and more efficient.

Goals and Research Questions

The inspiration for this research arises from prior work conducted at the AEON Lab at the University of Illinois Chicago, where spectral methods were first explored as a tool to understand the behavior of large neural networks. Building on this foundation, the present thesis investigates how spectral geometry and Random Matrix Theory can provide interpretable, mathematically grounded solutions to the pressing challenges of modern artificial intelligence. The need for such approaches is clear: large language models and other deep architectures are increasingly deployed in critical applications, yet they remain prone to hallucinations, brittle under distributional shifts, and prohibitively costly to scale. These limitations highlight the necessity of frameworks that not only improve performance but also enhance trustworthiness and efficiency in a principled manner.

The central goals of this thesis are therefore twofold. First, to develop methods that detect and anticipate failures in large language and vision-language models by analyzing the spectral dynamics of their hidden activations. Second, to design compression strategies that reduce the computational burden of deep networks without sacrificing accuracy, guided by rigorous statistical laws rather than heuristics. These goals naturally lead to the following research questions: Can spectral geometry serve as a compact and interpretable signal of reliability in large-scale models? And can Random Matrix Theory provide a principled foundation for identifying and retaining the causal structure necessary for efficient compression? Addressing these questions is the focus of this thesis, with the broader objective of advancing the reliability and sustainability of modern deep learning.

Thesis Structure

In Chapter 1 we provide the necessary background to this research, reviewing the theoretical foundations of spectral geometry, random matrix theory, and the fundamentals of deep neural networks and large language models. This chapter introduces the mathematical tools that will be applied throughout the thesis. Chapter 2 surveys related work, presenting prior approaches to hallucination detection, out-of-distribution recognition, and model compression. We compare black-box, grey-box, and white-box detection methods,

as well as traditional compression techniques such as pruning, low-rank approximations, and standard knowledge distillation. Chapter 3 introduces the first main contribution of this thesis, **EigenTrack**, and describes in detail its methodology, design, and experimental evaluation for hallucination and out-of-distribution detection in large language and vision-language models. Chapter 4 presents the second contribution, **RMT-KD**, and explains how Random Matrix Theory can be used to guide causal knowledge distillation for efficient network compression. This chapter also discusses its empirical results across natural language processing and computer vision benchmarks. Finally, Chapter 5 concludes the thesis by summarizing the key contributions, outlining limitations, and suggesting directions for future research at the intersection of spectral theory and deep learning.

1 | Background

In this chapter we review the theoretical and technological foundations of our work. We first outline modern deep learning architectures, focusing on large language models, convolutional networks, and vision-language systems. We then introduce the mathematical basis of spectral geometry and random matrix theory. Next, we discuss two central challenges in today's models: reliability, including hallucinations and out-of-distribution errors, and efficiency, with emphasis on compression and knowledge distillation. We also survey prior applications of spectral methods in vision and language models, and conclude with a unified perspective that links efficiency and reliability through spectral geometry.

1.1. Background: Foundations of Deep Learning

Machine Learning (ML) develops algorithms that learn patterns from data rather than following explicit rules. Classical ML methods such as decision trees, support vector machines, and k-nearest neighbors rely on handcrafted features and work well on structured data. Deep Learning (DL), a subfield of ML, uses deep neural networks to automatically learn hierarchical representations from raw inputs. Enabled by large datasets and GPU computing, DL has driven breakthroughs in computer vision, natural language processing, and multimodal tasks [29, 52].

Neural Networks and Representation Learning

Feed Forward Neural Networks (FNNs), Figure 1.1, apply weighted sums and nonlinear activations to compute the value for each neuron, e.g., $h^{(l)} = \sigma(W^{(l)}h^{(l-1)} + b^{(l)})$ [29]. Training with backpropagation allows them to approximate complex functions, while hidden layers learn representations that map raw data to informative latent features. Modern networks are often overparameterized and still generalize well [9, 114].

CNNs exploit spatial locality with convolution and pooling layers, Figure 1.2, producing hierarchical visual features with notorious architectures like AlexNet, VGG, and ResNet demonstrated scalability [34, 49]. A typical convolution filter computes $h^{(l)}_{i,j} = \sigma(\sum_{m,n} W^{(l)}_{m,n} x^{(l-1)}_{i+m,j+n} + b^{(l)})$, capturing local patterns that compose into higher-level fea-

tures. RNNs process sequences with a recurrent hidden state, e.g., $h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$, and gated variants such as GRU and LSTM mitigate long-range dependency issues using various level of memory gates.

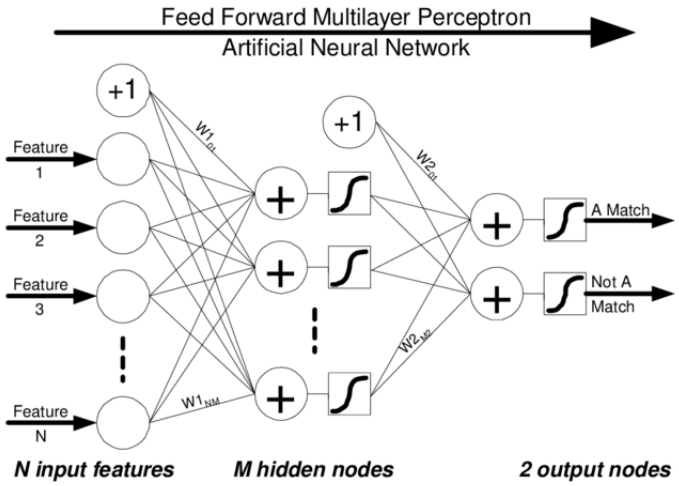

Figure 1.1: Feed Forward Neural network: features pass through hidden layers and non-linear activations, producing the final output [11].

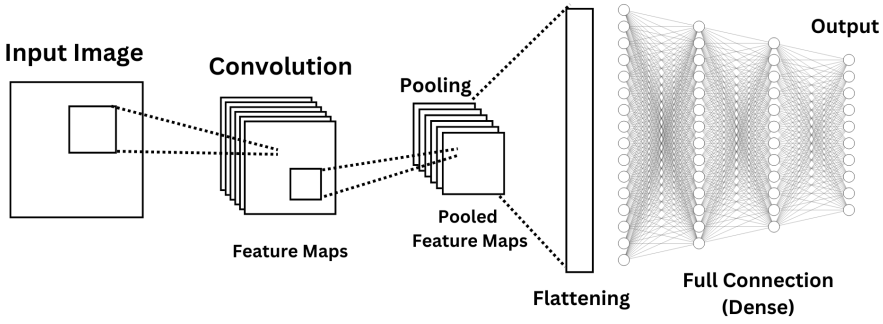

Figure 1.2: Convolutional Neural Network: convolutional filters extract features, pooling layers reduce resolution, and linear layers perform the final classification [15].

Transformers and Large Language Models

Transformers, introduced by Vaswani et al. [102], rely on self-attention to capture long-range dependencies without recurrence, Figure 1.3. The attention mechanism computes similarity scores between all pairs of tokens in a sequence, allowing each token to weight and aggregate information from every other token. This enables the model to focus on

relevant context regardless of distance. Their scalability has enabled Large Language Models (LLMs) such as BERT [18], GPT [80], and LLaMA [99]. This architecture has been shown to be effective also for multimodal data.

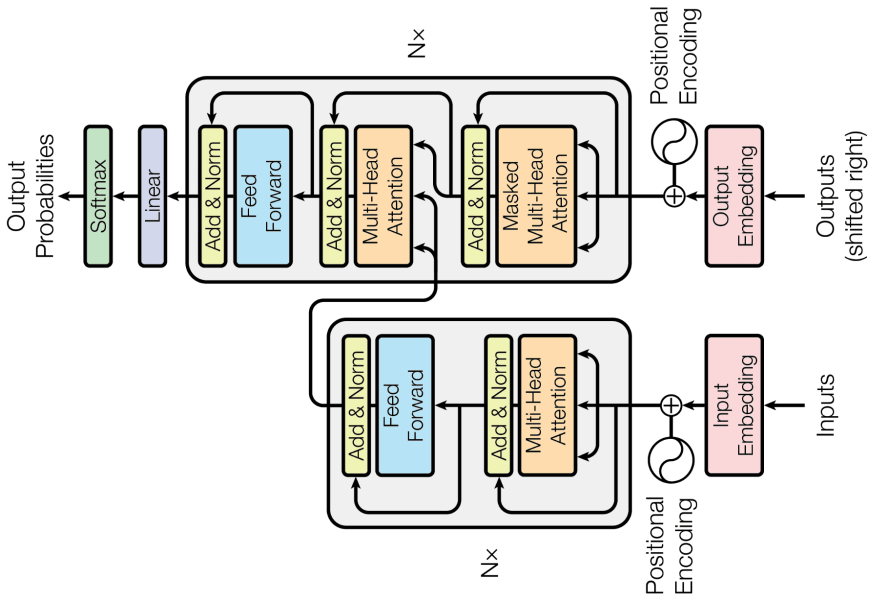

Figure 1.3: Transformer architecture [102]. Each encoder–decoder block uses multi-head self-attention followed by feed-forward layers and residual connections.

VLMs and overparameterization. VLMs align image and text modalities to enable captioning and visual QA; CLIP and LLaVA are prominent examples [61, 82]. Overparameterization enables interpolation yet increases cost; explaining why such models still generalize remains active, with spectral methods offering one perspective [9].

1.2. Background: Spectral Foundations

At the foundation of spectral geometry are **eigenvalues** and **eigenvectors**. For a square matrix $A \in \mathbb{R}^{n\times n}$, an eigenvector $v \neq 0$ and eigenvalue λ satisfy $Av = \lambda v$, identifying directions of pure scaling. Eigenvalues are roots of $\det(A - \lambda I) = 0$ [39]. In practice, QR, SVD, or iterative methods are used for large matrices [28].

For symmetric matrices, common in machine learning (e.g., covariance matrices), the spectral theorem guarantees a decomposition $A = Q\Lambda Q^T$ with orthogonal eigenvectors. The interpretation is that Eigenvalues quantify variance along eigenvector directions; dominant values indicate low-dimensional structure while a flatter spectrum suggests noise-like variability [100].

An important application is dimensionality reduction via PCA. After centering data $X \in \mathbb{R}^{n\times d}$, the covariance $C = \frac{1}{n}X^TX$ has eigenpairs (λ_i, v_i) and projecting onto the top-k eigenvectors yields $X_{red} = XV_k$ preserving the largest possible variance, Figure 1.4. In practice, SVD computes this efficiently without forming C explicitly.

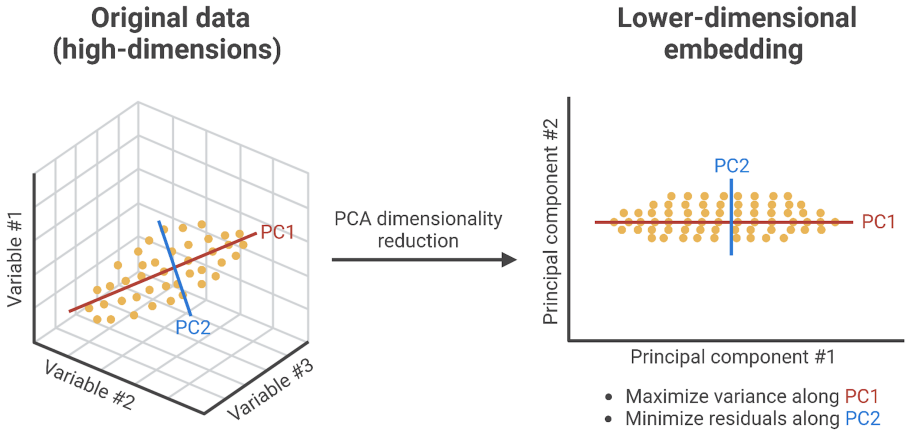

Figure 1.4: Principal component analysis: This image shows how Principal Component Analysis (PCA) reduces high-dimensional data into fewer dimensions by projecting it onto principal components [116]. PC1 captures the greatest variance, PC2 captures the largest remaining variance orthogonal to PC1, preserving essential structure with less complexity.

1.3. Background: Random Matrix Theory

Random Matrix Theory (RMT) studies the spectral properties of large matrices whose entries are random variables. Its central goal is to characterize the empirical spectral distribution (ESD), defined as the normalized histogram of eigenvalues of a random matrix as its dimensions grow. Let $\mathbf{M} \in \mathbb{R}^{p\times p}$ be a random symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$. The empirical spectral distribution is: $\mu_{\mathbf{M}}(\lambda) = \frac{1}{p}\sum_{i=1}^{p}\delta(\lambda - \lambda_i)$ where δ is the Dirac delta. As $p \to \infty$, $\mu_{\mathbf{M}}$ converges (in probability) to a deterministic limiting law, such as the Wigner semicircle or Marchenko–Pastur distribution, depending on the matrix ensemble.

Wigner semicircle law.

The Wigner Semicircle Law describes the eigenvalue density of large symmetric matrices with i.i.d. entries of mean zero and variance σ^2 [13]. Let $\mathbf{W} \in \mathbb{R}^{p\times p}$ with entries W_{ij} drawn i.i.d. and $\mathbf{W} = \mathbf{W}^T$. Then, as $p \to \infty$, the ESD converges to: $f_{\text{Wigner}}(\lambda) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - \lambda^2}$ for $|\lambda| \leq 2\sigma$, and 0 otherwise, Figure 1.5. This implies that the eigenvalues of a purely random symmetric matrix are asymptotically confined to the interval $[-2\sigma, 2\sigma]$, establishing a universal "noise floor."

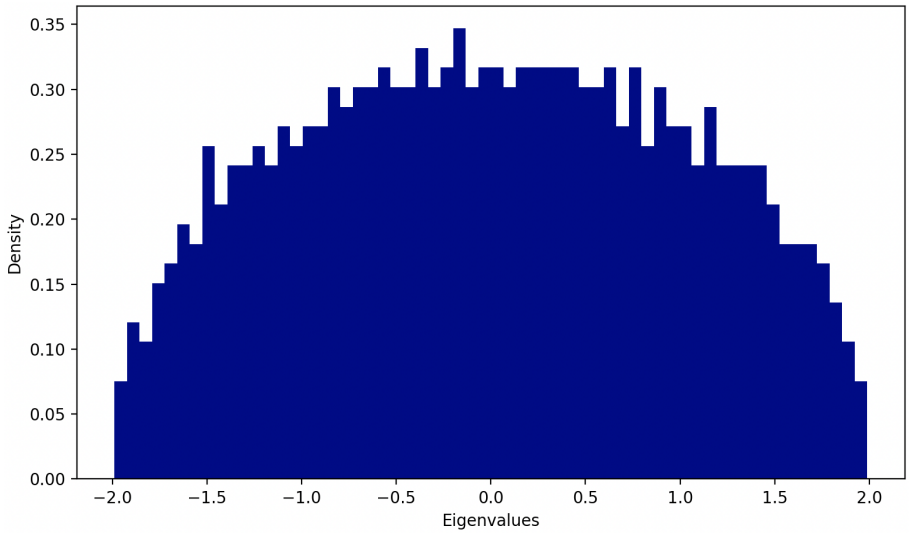

Figure 1.5: Wigner semicircle law: eigenvalue density of large symmetric random matrices.

Marchenko–Pastur distribution.

For covariance-type matrices, the limiting spectrum is described by the Marchenko–Pastur (MP) law [110]. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ have i.i.d. entries with mean zero and variance σ^2, and define the sample covariance matrix $\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$. If $n, p \to \infty$ with ratio $p/n \to c > 0$, then the ESD of $\mathbf{C}$ converges to: $f_{\text{MP}}(\lambda) = \frac{1}{2\pi c\sigma^2\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}, \quad \lambda \in [\lambda_-, \lambda_+]$ with support $\lambda_\pm = \sigma^2(1 \pm \sqrt{c})^2$, Figure 1.6. Thus, all eigenvalues are confined within $[\lambda_-, \lambda_+]$, forming the "bulk spectrum" of random covariance matrices.

Tracy–Widom distribution.

The Tracy–Widom distribution $F_\beta(s)$ describes the limiting fluctuations of the largest eigenvalue $\lambda_{\max}$ of large random matrices after proper centering and scaling, typically as $s = (\lambda_{\max} - \mu_n)/\sigma_n$, where μ_n is the spectral edge and $\sigma_n \sim n^{-2/3}$, which in case of the MP distribution is $\lambda_{\max} = \sigma^2(1 + \sqrt{c})^2 + \alpha n^{-2/3}X$ (where X follows the Tracy–Widom distribution and α is a costant factor). It arises universally for Gaussian matrices with index $\beta = 1, 2, 4$ corresponding to real symmetric (GOE), complex Hermitian (GUE), and quaternionic (GSE) cases. The distribution is defined through the Painlevé II equation $q''(x) = x\,q(x) + 2\,q(x)^3$ with boundary condition $q(x) \sim \text{Ai}(x)$ as $x \to +\infty$, where $\text{Ai}(x)$ is the Airy function. The cumulative form for $\beta = 2$ is $F_2(s) = \exp[-\int_s^\infty (x - s)\,q(x)^2\,dx]$, and for $\beta = 1$ it satisfies $F_1(s) = \sqrt{F_2(s)}\,\exp[-\frac{1}{2}\int_s^\infty q(x)\,dx]$, Figure 1.7. This law captures the universal edge behavior of eigenvalues in random matrix theory and appears in diverse high-dimensional systems.

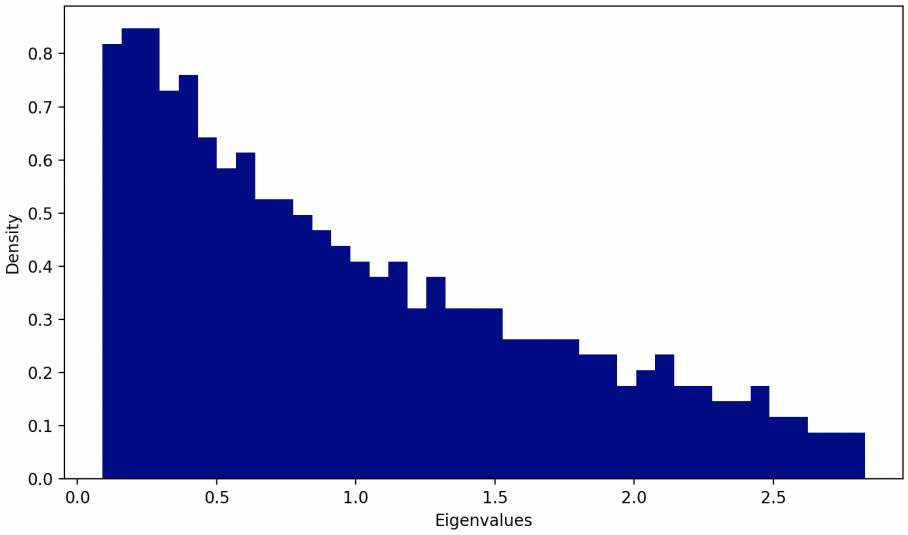

Figure 1.6: MP distribution: eigenvalue density of sample covariance (random) matrices.

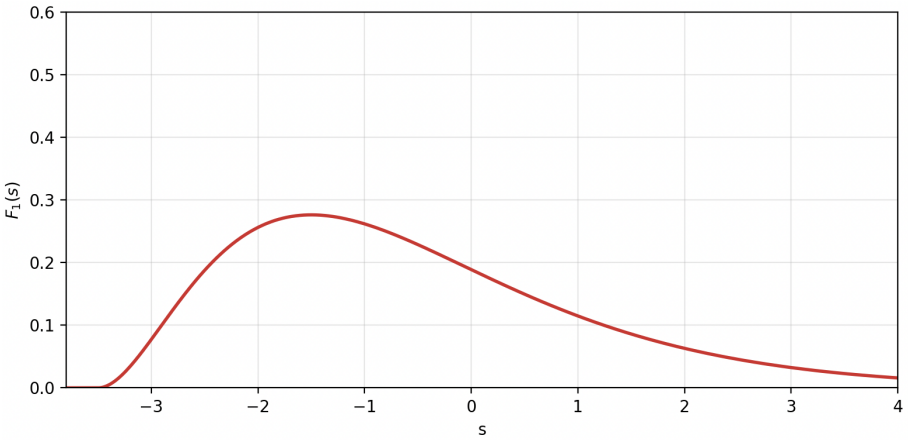

Figure 1.7: Tracy-Widom distribution: The variant with $\beta = 1$ (real numbers), describing the fluctuations of the largest eigenvalue of a random matrix around the spectral edge.

Spiked covariance model.

A fundamental extension of the MP model is the spiked covariance model [78], where a low-rank signal is embedded in isotropic noise. The population covariance is: $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p + \sum_{i=1}^{k} \theta_i u_i u_i^T$ where θ_i are spike strengths and u_i are orthonormal signal directions. We use Random Matrix Theory on the sample covariance matrices obtained from the LLMs activations on the datasets to identify the spikes components and the associated signal direction, which is useful for compression by projection and for comparing with the noise baseline.

BBP phase transition. The Baik–Ben Arous–Péché (BBP) phase transition [6] states that if the spike strength is below a critical threshold, the corresponding sample eigen-

values remain buried inside the MP bulk. Only when $\theta > \sigma^2(1+\sqrt{c})$ (equivalently, when population eigenvalues exceed $\lambda_+ = \sigma^2(1+\sqrt{c})^2$), do sample eigenvalues separate from the bulk and emerge as outliers: $\lambda_{\text{outlier}} > \lambda_+$. These detached eigenvalues carry information about the underlying signal, while eigenvalues within the MP bulk represent noise.

The BBP transition emerges directly from the upper edge λ_+ of the Marchenko-Pastur bulk by considering the asymptotic location of a sample eigenvalue $\hat{\lambda}$ stemming from an isolated population spike θ. This location is given by the mapping $\hat{\lambda}(\theta) \approx \sigma^2(\theta + \frac{c\theta}{\theta-1})$ for $\theta > 1$. The phase transition occurs precisely when this isolated eigenvalue detaches from the continuous bulk, i.e., when $\hat{\lambda}(\theta) = \lambda_+$. Substituting $\lambda_+ = \sigma^2(1+\sqrt{c})^2$ and solving the equation $\sigma^2(\theta + \frac{c\theta}{\theta-1}) = \sigma^2(1+\sqrt{c})^2$ for θ yields the critical threshold $\theta_{\text{BBP}} = \sigma^2(1+\sqrt{c})$. Thus, the BBP threshold is derived by finding the minimal population signal strength required to push a sample eigenvalue beyond the theoretical maximum λ_+ imposed by the noise.

1.4. Background: Hallucinations in AI Models

Large-scale neural networks have achieved remarkable success across natural language processing, computer vision, and multimodal reasoning. However, their reliability remains a central challenge. One of the most prominent issues is the phenomenon of *hallucination*, where a model generates fluent but factually incorrect or unsupported content. Hallucinations can emerge even in high-performing systems due to the mismatch between training distributions and real-world usage, or because models prioritize linguistic plausibility over factual grounding. This raises significant concerns in domains such as healthcare, law, and scientific discovery, where incorrect outputs may lead to serious consequences.

Closely related is the problem of *out-of-distribution* (OOD) generalization. Models are trained on large but finite corpora and may encounter inputs at inference that lie outside the training distribution. In these cases, internal representations can become unstable, often leading to either low-quality predictions or hallucinated outputs. This behavior illustrates a fundamental limitation: deep models tend to interpolate well within their training support but extrapolate poorly outside of it.

Uncertainty definition. Understanding these challenges requires considering the broader notion of *uncertainty* in machine learning. Uncertainty can be decomposed into two main types: *aleatoric* and *epistemic*. Aleatoric uncertainty arises from inherent noise or ambiguity in the data itself, such as ambiguous sentences or low-quality images. Epistemic uncertainty, on the other hand, reflects the model's lack of knowledge and tends to dominate when the model faces OOD inputs or insufficient training coverage. Formally, if

a model produces a predictive distribution $p(y|x, \theta)$ given parameters θ, aleatoric uncertainty is tied to the conditional variability in y for fixed θ, whereas epistemic uncertainty corresponds to variability across different plausible parameter settings. In practice, distinguishing between these two types is critical for reliability. For example, high aleatoric uncertainty may be unavoidable and can be communicated through probabilistic predictions, while epistemic uncertainty suggests the need for mechanisms such as abstention, retrieval, or further model adaptation. However, modern neural networks often underestimate their uncertainty because maximum-likelihood training encourages overconfident predictions. This miscalibration increases the likelihood that models generate hallucinations, as epistemic uncertainty is often underestimated in their predictive distributions. Such underestimation directly undermines reliability, especially in high-stakes applications.

1.5. Background: Compression in Deep Networks

The success of deep neural networks is partly due to their overparameterization, which provides strong representational capacity and generalization. However, this leads to high memory usage, inference latency, and energy costs. To mitigate these issues, several compression techniques have been proposed. In this section, we outline four main approaches: knowledge distillation, pruning, quantization, and sparsity. Pruning methods aim to remove redundant or unimportant parameters, typically weights or entire channels, from a trained model. The underlying assumption is that many connections in overparameterized networks contribute little to predictive performance. By eliminating these parameters, the model becomes smaller and faster while maintaining accuracy within acceptable bounds [32]. Quantization reduces the precision of weights and activations, replacing high-precision floating-point representations with lower-bit formats such as 8-bit integers. This decreases both the memory footprint and the cost of arithmetic operations, enabling deployment on resource-constrained devices. While aggressive quantization can cause accuracy degradation, careful design and post-training calibration mitigate this effect [27]. Sparsity-based approaches enforce or exploit structured or unstructured sparsity in network parameters. Unlike pruning, which is often applied after training, sparse methods may involve training models under sparsity constraints or using specialized architectures. Sparse representations reduce storage and allow faster computation on hardware that supports efficient sparse operations [24].

Knowledge distillation. Knowledge distillation transfers knowledge from a large teacher to a smaller student. The student is trained on labels and softened teacher outputs; with

teacher logits $z^{(T)}$ and student logits $z^{(S)}$, softened probabilities use temperature τ, and the objective combines cross-entropy with a KL term: $\mathcal{L} = \alpha\,\mathrm{CE}(y, p^{(S)}) + (1-\alpha)\,\tau^2\,\mathrm{KL}(p^{(T)} \,\|\, p^{(S)})$. This captures similarity structure in the teacher's predictions and yields compact models with competitive accuracy [38].

1.6. Background: Spectral Methods in Large Models

Spectral methods study how information in high-dimensional representations is distributed across directions or frequencies by inspecting the eigenstructure of matrices derived from data or model internals (e.g., activation covariances, Jacobians, or attention maps). In deep learning, these analyses help characterize global geometry (e.g., low-rank structure, anisotropy, and concentration), diagnose training dynamics and generalization, and suggest principled simplifications such as dimensionality reduction or subspace projections [68, 76].

Spectral lenses on neural representations. A common approach is to stack hidden activations across samples or time and examine covariance spectra. Eigenvalues describe variance concentration; rapid decay indicates low-rank structure, while heavy tails suggest richer features [68, 76]. SVCCA and CKA compare layer representations without supervision [48, 83].

Links to random matrix theory. When activations are approximately mean-zero and isotropic at a given scale, the empirical covariance spectrum exhibits a bulk consistent with random matrix predictions. The Marchenko–Pastur law describes the asymptotic eigenvalue density of sample covariance matrices with aspect ratio parameter and noise variance, providing a principled baseline for distinguishing noise-dominated directions from signal-bearing outliers; deviations from this baseline, such as outlier eigenvalues or widened gaps, indicate emergent structure [65]. For symmetric weight or Hessian-like operators with independent fluctuations, the Wigner semicircle distribution offers a corresponding null model [104]. In practice, deep networks rarely behave like pure noise, but these laws give calibrated reference points: directions with eigenvalues near the bulk edge are more likely to be correlational or redundant, while detached eigenvalues often track semantic or task-relevant factors, Figure 1.8.

Frequency and complexity biases. Spectral viewpoints illuminate which functions networks learn most readily. In overparameterized regimes, many models show a spectral bias toward low-frequency components before higher-frequency detail [84].

Applications to vision and language. Vision spectra often display heavy tails and class-conditional structure, while language models exhibit anisotropy and low-rank tendencies in embeddings and intermediate states [48, 68, 76, 83]. These analyses provide a unifying language for representation quality and compression.

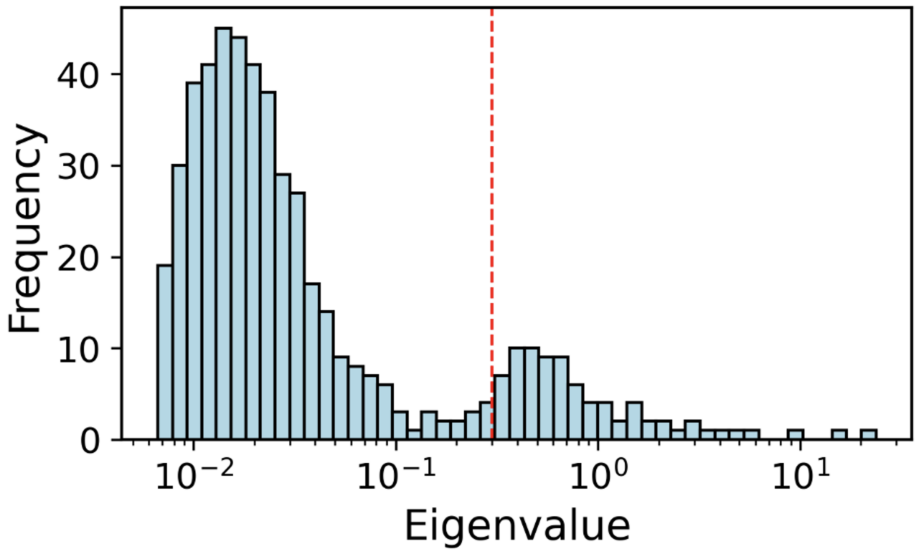

Figure 1.8: Empirical eigenvalue distribution: spectrum of hidden layer activations in a deep network, specifically the first layer of a BERT encoder during training. The histogram shows a bulk of small eigenvalues consistent with noise-like directions, while a few outliers (to the right of the red line) indicate informative, structured components. Such separation between bulk and outliers is the basis of spectral analyses in deep learning.

Implications for practice. Because spectral summaries are global and compact, they are useful for monitoring training health, identifying distributional shifts, and motivating architecture or optimization choices. For example, a pronounced low-rank structure motivates low-rank adapters and subspace updates, where model changes are confined to a small number of directions while preserving overall behavior [40]. More broadly, comparing empirical spectra to random matrix baselines provides quantitative signals for when representations become structured, when they remain noise-like, and how complexity evolves with data and compute.

1.7. Background: Efficiency and Reliability in AI

The rapid scaling of artificial intelligence (AI) systems has brought remarkable advances in natural language processing, computer vision, and multimodal learning. At the same time, this growth has exposed two central and often competing requirements: *efficiency*, the ability to deploy and operate models within practical computational budgets, and *re-*

liability, the assurance that predictions are robust, trustworthy, and aligned with human expectations. While these dimensions are frequently studied in isolation, they are deeply interconnected. Both can be analyzed through the lens of *spectral geometry*, which provides a principled framework for understanding the dynamics of high-dimensional neural representations.

Efficiency in large-scale AI models. Efficiency concerns the computational and resource costs of training and deploying large models, including latency, memory footprint, and energy [91]. Scaling laws indicate predictable performance gains with size and data [46], but at high financial and environmental cost. Compression and low-rank structure aim to remove redundancy without degrading accuracy [33].

Reliability in large-scale AI models. Reliability addresses trustworthiness of predictions. Large networks can hallucinate, exploit spurious correlations, and degrade under distribution shift [36, 74]. Standard confidence scores are often miscalibrated and miss deeper representational instabilities [30], making reliability critical in high-stakes domains.

Spectral geometry as a bridge. Although efficiency and reliability may seem distinct, both are tied to the geometry of neural representations. High-dimensional activations often concentrate around low-rank subspaces, and their covariance spectra reveal whether learned features are structured or dominated by noise [68]. Efficient models seek to eliminate redundant directions, while reliable models must ensure that remaining directions capture stable and causal signals. Spectral geometry provides the mathematical lens for unifying these perspectives: eigenvalue distributions and spectral entropy measure redundancy, while spectral gaps and deviations from random matrix baselines indicate stability and robustness. By analyzing activations through this shared framework, it becomes possible to design models that are both compact and trustworthy.

2 | Related Work

This chapter reviews prior work across hallucination detection, OOD detection, model compression, Random Matrix Theory in deep learning, and multimodal hallucinations. The goal is to highlight the strengths and limitations of existing approaches and to motivate the thesis framework: a unified spectral-geometry view in which the same eigenvalue structure supports reliability monitoring (EigenTrack) and principled compression (RMT-KD). The review is organized to mirror the thesis contributions: we first cover reliability-oriented methods, then efficiency-oriented compression, and finally the RMT literature that provides the mathematical bridge between the two.

2.1. Related: Hallucination Detection

Large language models exhibit "hallucinations": fluent statements that are unfaithful to sources or unverifiable by background knowledge. These failures undermine trust and safety, and persist even at scale because next–token prediction does not directly penalize confident fabrication [42, 59]. As a result, much work focuses on *detection* rather than elimination, spanning black–box, grey–box, white–box, and spectral families that trade off access, cost, and generality.

Black-box. Black–box detectors only observe input/output (optionally multiple samples). SelfCheckGPT measures agreement across re–generations using QA/NLI probes and reports strong zero–resource performance [64]. Other lines calibrate disagreement or use NLI chains (e.g., CoNLI) to flag unsupported claims [54, 101]. HaloScope leverages unlabeled generations to separate truthful and untruthful clusters, providing a practical detector without manual labels [20]. These methods are easy to retrofit and model-agnostic, but can be costly due to multiple generations and may fail on self–consistent errors [42, 59].

Grey-box and white-box. Grey–box detectors exploit limited internals such as token log–probabilities while avoiding access to hidden states. DetectGPT and Fast–DetectGPT use probability curvature to separate model outputs from human text [7, 71]; Glimpse reconstructs probability information to enable richer detectors on API-only models [8].

White–box detectors use internal activations or attention for real-time alarms: MIND trains a lightweight detector [92]; INSIDE derives an eigenvalue-based consistency score [14]; attention-pattern scoring links faithfulness to internal focus [90]. ReDeEP connects hallucination detection to mechanistic interpretability in RAG settings, highlighting how retrieval failures surface in internal signals [96]. These methods can be accurate and low-latency but require model internals and careful calibration, and may need re-tuning across model families [42, 101].

Spectral methods. Spectral detectors focus on eigenvalue structure rather than token probabilities. Under benign conditions, representations follow RMT bulk laws (e.g., MP), while failures induce outliers and widened gaps [5, 65]. Attention-based spectral methods compute Laplacian eigenvalues from attention graphs and use eigengaps as probes [10]; eigenvalue-based consistency scores assess dispersion of internal representations [14]. Spectral features are compact and interpretable and align well with the spectral-temporal approach of EigenTrack, but require careful control of windowing and layer choice [5, 65].

2.2. Related: Out-of-Distribution Detection

Modern deep networks often assume test data match training data; violations lead to the out-of-distribution (OOD) problem [109]. OOD detection flags such inputs to enable abstention or fallback, and in LLMs it often manifests as hallucinations or unstable performance under domain shifts.

Score-based methods. OOD detectors often rely on output statistics such as confidence or energy. MSP uses $\max_i p(y_i|x)$ [36], while ODIN adds temperature scaling and small perturbations to improve separation [58]. Mahalanobis scores model class-conditional features [53]. Energy-based scoring uses $E(x) = -T \log \sum_i e^{z_i/T}$ and is strong in both vision and LLMs [44, 62]. Sequence-level logit similarity and margin-based criteria have also been proposed for LLMs [106]. These methods are simple and post-hoc but can be sensitive to calibration, and they may conflate epistemic and aleatoric uncertainty in practice.

Representation-based methods. These methods leverage internal embeddings, using kNN or centroid distances and DkNN for layerwise conformity [75, 93, 97]. Subspace comparisons such as SVCCA/CKA reveal shifts in representation geometry, and attention entropy or hidden-state variance act as uncertainty signals in transformers [48, 83, 111]. They are interpretable and can localize failures to layers, but can be memory-intensive and sensitive to feature collapse or overfitting.

Spectral/statistical methods. Spectral approaches analyze covariance spectra and relate OOD to RMT: in-distribution data follow MP-like bulks, while OOD induces outliers or heavy tails. RankFeat uses effective rank [94], SpectralGap thresholds eigengaps [77], and SNoJoE tracks singular-value energy across layers [2]. These methods are unsupervised and architecture-agnostic, align naturally with spectral monitoring in EigenTrack, but can be sensitive to batch size and numerical precision; randomized eigensolvers and moving-window estimates mitigate cost [19, 67].

Across benchmarks, score-based methods are often competitive but degrade under severe distribution shift or label-shifted test sets, while representation-based and spectral methods tend to be more stable when internal features remain well-structured. In practice, hybrid systems that combine a fast score-based trigger with a slower representation or spectral check can improve robustness while maintaining low latency. This motivates EigenTrack's design, which retains a lightweight computation path but incorporates spectral structure for stability across shifts.

2.3. Related: Model Compression

Deploying modern networks under latency and energy constraints motivates compression. Four families dominate: pruning, quantization, low-rank factorization, and knowledge distillation.

Pruning. Pruning removes parameters deemed unnecessary. Unstructured magnitude pruning can reach high sparsity with minimal loss, as in deep compression [31], while lottery ticket results suggest dense networks contain sparse subnetworks that train well [25]. More precise criteria use gradients or second-order information (e.g., Taylor or WoodFisher) to estimate loss impact [72, 89]. Structured pruning removes channels, heads, or blocks to obtain hardware-friendly speedups; network slimming and attention head pruning are representative examples [63, 70]. These methods can yield real speed gains but require careful tuning and kernel support, and they often trade sparsity for irregular memory access on GPUs.

Quantization. Quantization maps weights and activations to low precision. Post-training quantization is simple but can struggle with activation outliers in LLMs; SmoothQuant and GPTQ mitigate this with calibration or second-order updates [26, 107]. Quantization-aware training improves low-bit accuracy at the cost of additional fine-tuning [21]. Practical toolchains include per-channel affine quantization and mixed-precision schemes [41]. Realized speedups depend on kernel support and operator coverage, and aggressive quantization can degrade calibration.

Low-rank factorization. Low-rank factorization replaces large matrices with products of smaller ones, reducing parameters and compute. SVD-based decompositions work well for fully connected layers [17], while tensor decompositions apply to convolutions [51]. In transformers, ALBERT and LoRA highlight low intrinsic dimensionality and parameter sharing [40, 50]. These methods are dense and hardware-friendly but require choosing ranks per layer and can underfit if ranks are too small; intrinsic-dimension analyses provide guidance on feasible rank budgets [1].

Knowledge distillation. Distillation trains a compact student to match a larger teacher's outputs, combining task loss with KL divergence on soft targets [37]. Variants include FitNets, attention transfer, and contrastive distillation [86, 98, 113]. In NLP, DistilBERT, PKD, Theseus, and MiniLM show strong compression for transformers [88, 95, 103, 108]. Once-for-All style supernet training can distill multiple subnetworks with varying widths/depths [12]. Distillation is flexible and pairs well with pruning/quantization, but it does not explicitly identify which representation directions are causal, which motivates RMT-KD.

In practice, compression pipelines often combine families (e.g., pruning + quantization + distillation) to recover accuracy after structural changes. This highlights a gap: most methods optimize efficiency but do not explicitly use representation geometry to decide what to keep. RMT-KD addresses this by selecting directions based on statistically significant eigenvalue structure, making the compression criterion data-driven rather than purely heuristic.

2.4. Related: Random Matrix Theory in DL

Random Matrix Theory (RMT) provides a statistical–mechanical lens to analyze deep neural networks, treating weight and activation matrices as high-dimensional random systems. By studying the eigenvalue spectra of these matrices, RMT reveals universal behaviors that shed light on learning dynamics, generalization, and robustness. Unlike heuristic diagnostics, spectral geometry provides analytical tools to quantify phase transitions, implicit regularization, and signal–noise separation in modern networks. This section summarizes key developments applying RMT to deep learning and related areas of reliability and robustness.

Phase transitions. A central insight from RMT is that eigenvalue distributions of random covariance matrices follow the Marchenko–Pastur (MP) law, which describes the limiting density of eigenvalues for a matrix $XX^\top/n$ when both n and the feature dimension grow large [65]. In deep networks, deviations from this theoretical bulk indicate

structured correlations in learned representations. When training begins, the spectrum typically resembles the MP bulk; as learning progresses, a few eigenvalues detach from the bulk, forming outliers associated with meaningful features or "spikes." This transition is governed by the Baik–Ben Arous–Péché (BBP) threshold [5], which determines when a signal eigenvalue separates from random noise.

Empirical studies demonstrate that layer weight spectra in trained networks display heavy-tailed distributions following power laws rather than pure MP shapes [66, 67]. These heavy tails reflect correlations induced by gradient descent and act as fingerprints of generalization. The phase-transition view thus links RMT to implicit bias: well-generalizing models operate near criticality, balancing order (signal) and chaos (noise). Conversely, overtrained or overregularized models exhibit either degenerate (collapsed) spectra or overly flat ones, signaling underfitting or overparameterization. This perspective motivates using bulk–outlier separation as a principled selection rule in compression and reliability monitoring.

Implicit regularization. RMT also explains the implicit regularization mechanisms that emerge in large-scale optimization. Gradient descent, batch normalization, and dropout collectively push weight matrices toward critical spectral regimes, shaping their eigenvalue decay [67]. This self-organized criticality aligns with the notion of *spectral bias*, where networks learn low-frequency, smooth components first. In the spectral domain, weight and gradient covariance spectra often show $1/\lambda^{\alpha}$ power-law tails with α between 2 and 5, characterizing an intermediate regime between random noise and strong correlations.

Such spectral regularization correlates with model robustness and generalization: networks with heavy-tailed spectra resist overfitting and exhibit better OOD performance [69, 79]. Moreover, the Hessian spectra of well-trained models typically have a small number of large eigenvalues corresponding to informative directions and a bulk near zero corresponding to flat minima. This separation implies that stochastic optimization implicitly enforces a spectral cutoff, analogous to Tikhonov regularization in inverse problems. In this view, spectral analysis provides both a diagnostic and a theoretical justification for why simple optimization procedures produce well-behaved models without explicit constraints.

Subspace analysis. Beyond bulk statistics, RMT provides tools for subspace analysis of learned representations. In deep networks, activations at each layer can be viewed as samples from a high-dimensional manifold embedded in feature space. The covariance spectrum captures how variance is distributed across latent directions. Signal–noise decomposition based on spectral truncation isolates meaningful subspaces from isotropic

noise, yielding insights into feature redundancy and causal directions. Studies show that informative subspaces correspond to outlier eigenvectors, while the noise subspace follows MP-like statistics [67, 87]. Monitoring spectral entropy or eigengaps across layers reveals how information condenses as depth increases, forming hierarchical subspaces of decreasing intrinsic dimension.

Another active thread connects RMT to training dynamics and scaling. As models grow, spectra often become more structured, with clearer outlier separation and heavier tails, suggesting that increased capacity can amplify meaningful low-dimensional structure rather than only adding noise [66, 67]. This has practical implications for compression: larger models may admit more aggressive dimensionality reduction while retaining accuracy, because their representation subspaces are richer but more redundant. RMT-inspired diagnostics have also been used to assess layer health, detect over-regularization, and compare optimization regimes by tracking how spectra evolve during training.

Recent work extends subspace spectral analysis to multimodal models. In vision–language transformers, joint embeddings exhibit coupled spectral dynamics; alignment across modalities manifests as correlated eigenvalue shifts between text and vision subspaces. Spectral filtering, where activations are projected onto dominant eigenspaces, has been shown to enhance robustness and interpretability. In particular, *EigenShield* introduces causal subspace filtering via RMT principles to defend against adversarial perturbations in vision–language models, identifying and suppressing anomalous eigenmodes while preserving semantic structure [16]. This connects RMT-based subspace analysis to adversarial security, suggesting that spectral geometry can serve both analytical and defensive roles.

2.5. Related: Multimodal Models and Hallucination

The integration of visual, textual, and sometimes auditory modalities has enabled a new generation of foundation models capable of perception and reasoning across heterogeneous data. Vision–language models (VLMs) such as CLIP, BLIP, Flamingo, LLaVA, and GPT-4V have demonstrated impressive zero-shot capabilities on diverse multimodal tasks including captioning, VQA, and visual reasoning [3, 56, 60, 73, 81]. However, the same scaling that endows such flexibility also amplifies problems of reliability. Multimodal hallucination, the generation of textual or visual content inconsistent with the actual input image or video, emerges as a major limitation [4]. These hallucinations can manifest as fabricated objects in image captions, incorrect visual attributes, or semantically inconsistent reasoning steps, threatening the deployment of VLMs in sensitive contexts such as medical imaging, autonomous driving, or assistive technology.

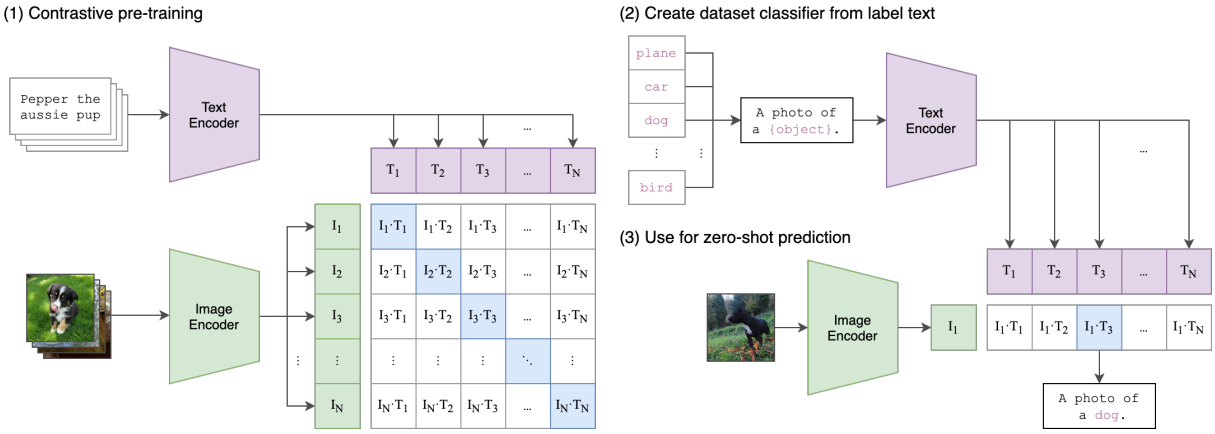

Figure 2.1: CLIP schematic: CLIP aligns images and text by projecting them into a shared latent space [82], espcially powerful fror zero-shot predictions.

Nature of multimodal hallucinations. Multimodal hallucination differs from text-only hallucination because it stems from misalignment between modalities. Empirical studies classify hallucinations into *object*, *attribute*, and *contextual* types [4, 85]. Causes include dataset biases, imbalance between vision and language encoders, and decoding that privileges linguistic priors over perceptual grounding.

Early VLMs such as OSCAR and VinVL relied heavily on object detection features, which constrained hallucination but limited generality [57, 115]. End-to-end pretrained architectures like CLIP and BLIP2 improved generalization but exhibited increased hallucination due to weaker explicit grounding, Figure 2.1. Recent works have shown that model scale and instruction-tuning can paradoxically increase hallucination frequency even while improving benchmark accuracy, as seen in GPT-4V and other multimodal LLMs [60, 73]. This tension highlights a key challenge: multimodal grounding requires both accurate fusion and balanced modality dominance.

Detection and mitigation strategies. Detection adapts linguistic and perceptual uncertainty measures. Black-box metrics evaluate semantic alignment between captions and visual regions; VisualHallucination provides benchmarks and metrics [4]. Grey-box approaches use attention entropy or cross-modal similarity matrices [112]. White-box or spectral detectors analyze fusion-layer covariance spectra, where deviations or gaps signal misalignment consistent with RMT-based diagnostics.

Recent work also explores retrieval-augmented and grounding-augmented pipelines, where external visual evidence or region tags are injected to constrain generation. These methods can reduce hallucinations but add latency and require careful alignment between retrieved context and the model's internal representations. In practice, multimodal systems often

combine lightweight confidence checks with heavier post-hoc verification, reflecting the same cost–accuracy trade-off seen in text-only settings.

Mitigation techniques span architectural, training, and decoding interventions. Architecturally, contrastive pretraining with balanced image–text pairs (e.g., CLIP, ALIGN) reduces dataset-induced biases [43, 81]. Training-level solutions such as grounding instruction tuning explicitly penalize ungrounded generations by augmenting prompts with object tags or region features [55, 60]. Reinforcement learning with human feedback (RLHF) has also been adapted to multimodal settings, optimizing for human-judged visual faithfulness instead of text coherence [73]. Decoding-level methods, including constrained beam search or re-ranking by cross-modal consistency, filter out hallucinated outputs post hoc [85].

Spectral and causal perspectives. Multimodal fusion layers couple visual and textual subspaces; well-grounded models show smooth decay in the joint spectrum, while hallucinations correlate with spurious outliers and eigengaps [16, 67]. Spectral regularization and subspace filtering can mitigate hallucinations by projecting onto stable causal eigenspaces. RMT-informed analysis also aids interpretability by isolating which eigenvectors carry cross-modal signal versus noise.

3 | EigenTrack: Spectral Detection Framework

This chapter is based in part on the author's publication "EigenTrack: Temporal Spectral Analysis of Hidden Activations for Hallucination and OOD Detection" [23]. The text and figures have been expanded and reformulated for clarity and completeness. Paper available on arXiv and submitted to IJCNN 2026 conference, under review. See Appendix A

3.1. Methodology

EigenTrack introduces a real-time, interpretable framework for detecting hallucinations and out-of-distribution (OOD) behavior in large language and vision–language models. Rather than relying on surface-level probabilities, EigenTrack monitors how the internal geometry of activations evolves during generation, Figure 3.1. The key idea is that when a model begins to hallucinate or drift from the training distribution, its internal representations lose structure and become increasingly isotropic. This degradation can be quantified through the eigenvalue spectrum of hidden-layer covariances.

Hallucinations in large generative models often emerge gradually, not as isolated anomalies. During early steps of generation, activations are structured and highly correlated, reflecting a coherent internal reasoning process. As hallucination begins, these correlations decay and the internal state becomes more random. Random Matrix Theory (RMT) provides a natural lens for analyzing this behavior. In- distribution activations produce covariance spectra with a few large, dominant eigenvalues. When representations lose structure, the spectrum collapses toward the Marchenko–Pastur (MP) law, which characterizes uncorrelated random noise. By tracking spectral statistics over time, one can therefore detect when the model's reasoning dynamics begin to diverge from stable behavior.

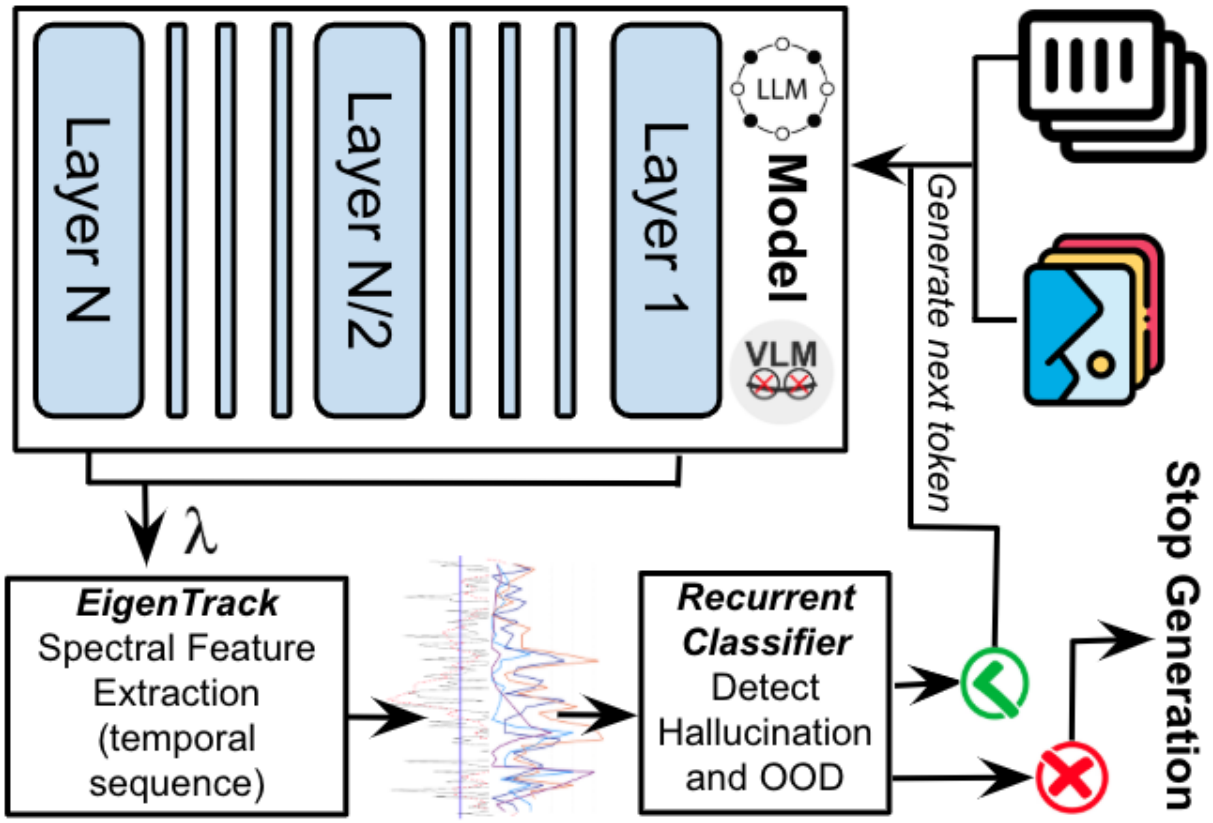

Figure 3.1: General EigenTrack architecture: Spectral features extracted from hidden activations are streamed into a recurrent discrepancy detector, which outputs early warnings of hallucination or OOD drift. In this pipeline, the problem is framed as binary classification.

Overview of the Framework

EigenTrack operates as an auxiliary monitoring head attached to any pretrained model, requiring no modification of model parameters. The pipeline comprises three components:

- **Spectral Feature Extraction:** Hidden activations from selected layers are collected in a sliding temporal window. Their covariance matrices are analyzed through a truncated singular value decomposition (SVD) to obtain eigenvalues describing the structure of the hidden-state manifold.
- **Spectral Discrepancy Detection:** From each window, a compact vector of spectral descriptors is computed, including entropy, leading eigenvalues, spectral gaps, and divergence from the MP distribution. These descriptors reflect how structured or random the current representations are.
- **Temporal Modeling and Early Warning:** A lightweight recurrent network ob-

serves the evolution of spectral descriptors over time. By learning patterns associated with representational collapse, it outputs a probability that the model is entering a hallucinatory or OOD regime.

Spectral Feature Extraction from Hidden Representations

Let the model layers be $L_1, L_2, \ldots, L_m$. At each generation step t, activations from the selected layers are concatenated: $v_t = [h_{1,t} \| h_{2,t} \| \cdots \| h_{m,t}], \quad h_{\ell,t} \in \mathbb{R}^d$. To capture short-term temporal evolution, EigenTrack maintains a sliding window of the most recent N steps: $H_t = \left[v_{t-N+1}, \ldots, v_t\right]^\top \in \mathbb{R}^{N \times md}$. The empirical covariance is then $C_t = \frac{1}{N} H_t^\top H_t$ and its eigenvalues are obtained efficiently via the truncated singular value decomposition $H_t = U_t \Sigma_t V_t^\top, \quad \lambda_{t,i} = \frac{\sigma_{t,i}^2}{N}$ Because $N \ll md$, the SVD is computationally cheap, and randomized or incremental updates allow near real-time operation during autoregressive decoding, Figure 3.2. Each $\lambda_{t,i}$ measures how variance is distributed across activation directions, providing a geometric fingerprint of the model's internal dynamics.

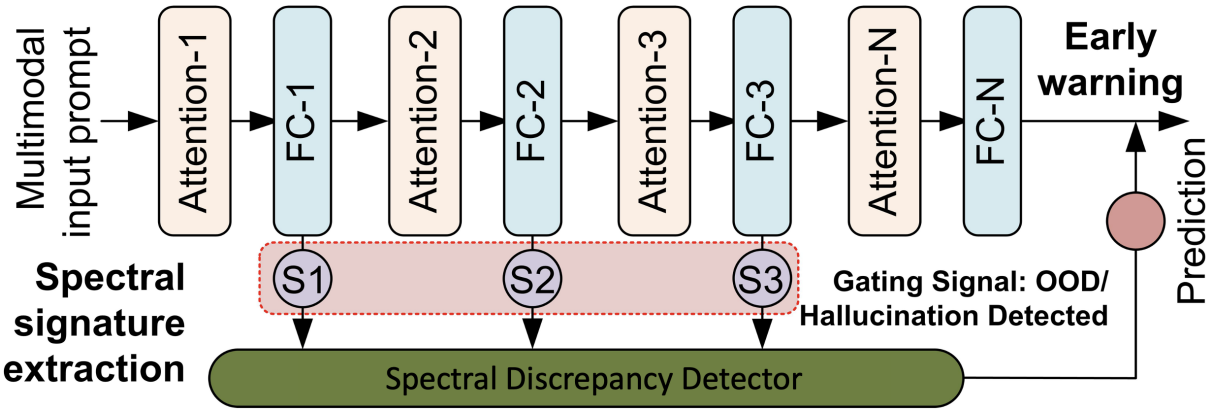

Figure 3.2: Layer-level EigenTrack feature extraction: Selected layers feed activations into the spectral feature extractor, which computes eigenvalue-based descriptors and streams them to the discrepancy detector in real time.

Spectral Descriptors

From the spectrum $\{\lambda_{t,i}\}$, EigenTrack derives a compact, evolving vector F_t of interpretable descriptors, Figure 3.3. It includes 22 features in total, among which we have:

- **Leading Eigenvalue Sum:**

$$s_t = \sum_{i=1}^{5} \lambda_{t,i} \tag{3.1}$$

  measuring how much variance is concentrated in the top principal directions. Larger values indicate low-rank, structured representations, while smaller values suggest diffuse or noisy activations.

- **Spectral Entropy:**

$$S_t = -\sum_i p_{t,i} \log p_{t,i}, \qquad p_{t,i} = \frac{\lambda_{t,i}}{\sum_j \lambda_{t,j}} \tag{3.2}$$

  where high entropy indicates isotropic, less informative embeddings, and low entropy reflects concentrated, coherent feature directions.

- **KL Divergence from the Marchenko–Pastur Baseline:**

$$D_{KL}\big(p(\lambda_t) \,\|\, \rho_{MP}(\lambda)\big) \tag{3.3}$$

  where $\rho_{MP}(\lambda)$ denotes the Marchenko–Pastur eigenvalue density. Small divergence suggests noise-like, unstructured activations; larger divergence indicates meaningful statistical structure. The same also applies when quantifying the difference in distributions with the **Wasserstein distance**, which is another feature we consider in our classifier.

- **Tracy–Widom Fluctuation of the Leading Eigenvalue:** The probability that $\lambda_{t,1}$ significantly deviates from the MP bulk edge λ_+:

$$\mathbb{P}\big(\lambda_{t,1} > \lambda_+ + \delta\big) \tag{3.4}$$

  A substantial deviation indicates emergence of a dominant, non-random direction (signal-bearing subspace).

- **Spectral Gap Ratio:**

$$g_{t,i} = \frac{\lambda_{t,i}}{\lambda_{t,i+1}}, \qquad i \in \{1, \ldots, k\} \tag{3.5}$$

capturing how sharply leading eigenvalues detach from the rest. Larger gaps mark transitions between structured and noise-dominated subspaces.

- **Spectral Skewness:**

$$\gamma_t = \frac{1}{d}\sum_{i=1}^{d}\left(\frac{\lambda_{t,i} - \bar{\lambda}_t}{\sigma_t}\right)^3 \tag{3.6}$$

Positive skewness implies a heavy tail of large eigenvalues (structured information), while near-zero skewness suggests noise-like balance.

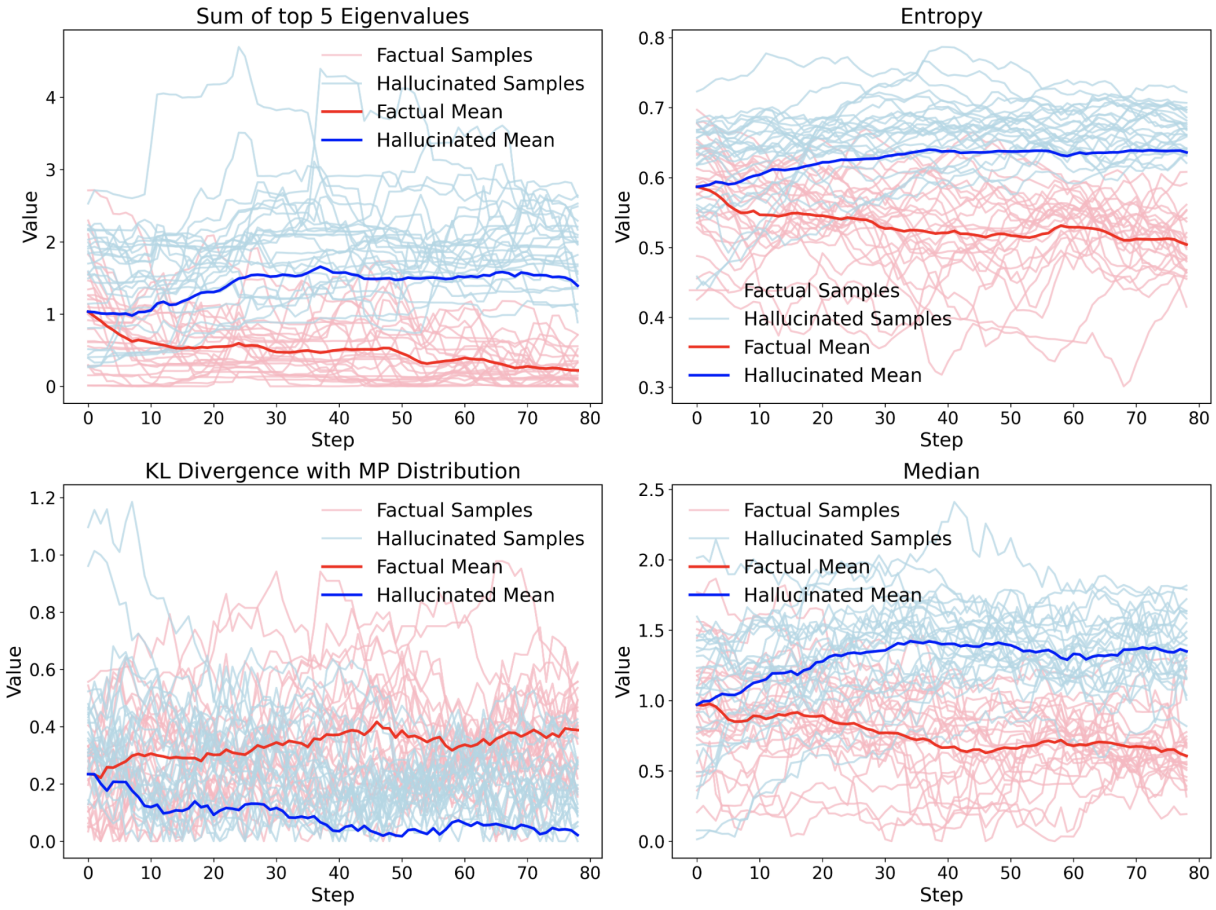

Figure 3.3: Temporal evolution of spectral statistics: The plots illustrate how spectral features evolve over time. Hallucinated sequences concentrate variance in a few dominant directions, producing higher cumulative sums of the top eigenvalues (top-left). Their spectra are flatter and more dispersed, resulting in elevated entropy values (top-right). From a random matrix theory perspective, hallucinated sequences remain closer to the noise-like regime, yielding lower KL divergence values (bottom-left), while factual sequences diverge more strongly, consistent with structured and informative dynamics. The median eigenvalue (bottom-right) is higher and more stable for hallucinations, whereas factual sequences show lower and more variable medians that correlate with model confidence.

Temporal Modeling with Recurrent Tracking

A single spectral snapshot can reveal instability, but hallucinations typically evolve gradually as the internal representation quality degrades over time. To capture this temporal evolution, EigenTrack treats the descriptors as a time series $\{F_1, F_2, \ldots, F_T\}$ and processes them using a small recurrent neural network. The feature sequence $(F_1, \ldots, F_T)$ is modeled as a multivariate time series and processed by a lightweight recurrent model (RNN, GRU, or LSTM). At each step, F_t enters the recurrent cell, hidden states propagate contextual information, and a feed-forward head outputs a binary logit corresponding to in-distribution versus anomalous context. Weight sharing across time ensures that the parameter count remains independent of sequence length, allowing the model to learn characteristic spectral signatures associated with stable or unstable dynamics.

This recurrent formulation provides three main advantages. First, it preserves temporal continuity by maintaining a memory of previous spectral states, enabling the detection of gradual drifts in representation quality. Second, it is computationally efficient, with updates occurring in constant time per step and minimal additional overhead. Third, it preserves causality by processing only past information, making it suitable for real-time inference and early warning during generation. At each generation step, the hidden state evolves according to $z_t = \mathrm{RNN}(F_t, z_{t-1})$, and the anomaly probability is computed as $\hat{y}_t = \sigma(W z_t + b)$. If $\hat{y}_t > \tau$, the detector raises a gating signal that can intervene by halting decoding, triggering retrieval grounding, or lowering the generation temperature. The recurrent head is highly compact, containing only a few thousand parameters, and can operate concurrently with model inference without significant latency.

EigenTrack supports tuning of the monitored layer set L, window length N, number of spectral features k, and recurrent hidden size to balance accuracy and latency. The pipeline also extends naturally to multimodal architectures by constructing H_t from cross-modal fusion layers or vision encoder blocks, enabling spectral monitoring across both language and visual representations within a unified latent space.

Efficiency and Practical Considerations

EigenTrack is designed to be lightweight, modular, and fast enough to operate in real time alongside large generative models. In practice, only a small subset of layers is monitored, typically one every three or four transformer blocks, since adjacent layers exhibit highly correlated spectral behavior. This sampling strategy minimizes data transfer without sacrificing sensitivity. Covariance estimation is implemented through truncated or incremental SVD, which avoids full matrix reconstruction and enables efficient eigenvalue

updates at each decoding step. The recurrent head itself contains only a few thousand parameters and performs constant-time updates, making it several orders of magnitude smaller and faster than transformer-based temporal models. For this reason, recurrent networks were preferred over transformers: they preserve causal processing, require negligible memory, and operate with minimal computational overhead. As a result, EigenTrack can be integrated directly into autoregressive decoding loops, providing continuous reliability monitoring with only a small additional latency.

3.2. Theoretical Justification

Spectral Geometry as a Diagnostic Lens

The theoretical foundations of EigenTrack arise from the observation that large neural networks, despite their nonlinear nature, exhibit emergent regularities that can be studied through the geometry of their hidden representations. The key insight is that during normal operation, the network's representations inhabit a structured, low-dimensional manifold, while under distributional shift or hallucination-inducing conditions, this manifold collapses toward isotropic randomness.

Modern decoder-only LLMs apply LayerNorm at every transformer block [80], centering activations and scaling them to unit variance. Combined with the near-orthogonality of projection matrices induced by weight decay and adaptive optimization methods such as Adam [47], this makes per-token activations approximately mean-zero and isotropic. Under these assumptions, the activations of each layer can be modeled as random samples from an approximately isotropic Gaussian ensemble, and their covariance spectra follow well-known laws from Random Matrix Theory (RMT). Departures from this baseline indicate the emergence of structure or instability within the model's internal representations. When the network encounters out-of-distribution (OOD) inputs or hallucination-inducing contexts, hidden activations exhibit correlated, low-rank perturbations that follow the spiked covariance model.

The Random Matrix Baseline

Random Matrix Theory provides analytical tools to model the spectrum of large, unstructured covariance matrices. Consider a matrix $X \in \mathbb{R}^{N \times d}$ whose entries are independent and identically distributed with zero mean and variance σ^2. In the asymptotic limit where $N, d \to \infty$ and the aspect ratio $q = d/N$ is fixed, the empirical distribution of eigenvalues of the sample covariance $C = \frac{1}{N} X^\top X$ converges to the *Marchenko–Pastur (MP) law* [65].

The density of eigenvalues under this null model is given by

$$\rho_{\mathrm{MP}}(\lambda) = \frac{1}{2\pi\sigma^2 q\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}, \quad \lambda \in [\lambda_-, \lambda_+] \tag{3.7}$$

with the support edges defined as $\lambda_\pm = \sigma^2(1 \pm \sqrt{q})^2$ This distribution characterizes the "noise floor" of high-dimensional representations. Any activation covariance spectrum that aligns closely with ρ_{MP} can be interpreted as statistically equivalent to a random ensemble with no significant correlations. Conversely, systematic deviations, such as the emergence of outlier eigenvalues or heavy tails, signal the existence of structured correlations and dependencies in the underlying representations. A similar concept, related to the top eigenvalue deviation, is modeled by the Tracy-Widom distribution.

An analogous principle holds for uncentered activations or correlation matrices whose entries are symmetrized rather than formed from $X^\top X$. In this setting, the eigenvalue density converges to the *Wigner semicircle law* [105], which describes the distribution of eigenvalues of symmetric random matrices. The Marchenko–Pastur, Tracy-Widom and Wigner laws form the theoretical backbone of EigenTrack's null model: they represent the spectral geometry of complete randomness against which meaningful structure can be contrasted.

Signal Emergence on the Spiked Covariance Model

The empirical covariance of a trained model's activations rarely conforms to the pure MP law. Instead, it follows a *spiked covariance model* [45], in which a low-rank signal subspace is embedded within a high-dimensional noisy background: $C_t = \sigma^2 I + \sum_{i=1}^{k} \theta_i u_i u_i^\top$ Here, $\sigma^2 I$ represents isotropic noise, and $\{u_i\}_{i=1}^{k}$ are orthogonal directions capturing coherent, semantically meaningful structure. The scalars θ_i denote the signal strengths along those directions. In the spectral domain, the presence of such low-rank perturbations produces outlier eigenvalues that detach from the MP bulk. This phenomenon is governed by the *Baik–Ben Arous–Péché (BBP) phase transition* [6], which defines a critical signal-to-noise threshold:

$$\lambda_s > \sigma^2(1 + \sqrt{q}) \quad \Longrightarrow \quad \text{signal eigenvalue separates from the bulk.} \tag{3.8}$$

When λ_s falls below this threshold, the signal direction becomes indistinguishable from noise, and its corresponding eigenvalue reabsorbs into the random bulk. This threshold delineates the boundary between ordered and disordered regimes in the representation space. Within deep neural networks, such transitions correspond to phases where la-

tent representations lose alignment with meaningful semantic directions—an effect that empirically correlates with hallucinations or failures of grounding in LLMs and VLMs.

From Static Spectra to Temporal Dynamics

Traditional RMT analyses treat C_t as a static object, characterizing one snapshot of representational geometry. However, in autoregressive models, activations evolve dynamically with each generated token. Let $\{\rho_t(\lambda)\}_{t=1}^{T}$ denote the sequence of spectral densities observed across a generation trace. EigenTrack extends the RMT framework by modeling this sequence as a stochastic process in the space of spectral measures. This approach captures not only instantaneous anomalies but also the temporal precursors of instability.

Formally, define a set of spectral statistics $\phi_t = f(\rho_t)$ summarizing the current spectral state (e.g., curvature, skewness, or divergence from MP baseline). The trajectory $\{\phi_t\}$ can then be regarded as a dynamical system: $\phi_{t+1} = \mathcal{F}(\phi_t) + \epsilon_t$ where $\mathcal{F}$ denotes the underlying transition operator induced by the model's internal dynamics and ϵ_t is a stochastic perturbation. When the model processes coherent, in-distribution input, $\mathcal{F}$ is approximately stationary: the geometry of the activation manifold evolves smoothly. Under hallucination or distributional shift, however, the transition dynamics become unstable, producing abrupt spectral fluctuations and non-stationarity in ϕ_t. These deviations serve as early-warning signals of representational collapse.

Implications for Model Reliability

In summary, the theoretical justification of EigenTrack rests on three pillars:

1. **Random Matrix Theory:** establishes the null hypothesis of isotropic randomness, represented by the Marchenko–Pastur and Tracy–Widom distribution.
2. **Spiked Covariance Theory:** explains how deviations from the MP law arise from coherent, task-aligned structure, and how their disappearance signals representational collapse.
3. **Temporal Spectral Dynamics:** connects the evolution of eigenvalue distributions to the stability of the model's internal manifold during generation.

Together, these principles provide a rigorous mathematical framework linking spectral geometry to reliability in deep generative models. EigenTrack operationalizes this connection by tracking the time evolution of the spectrum, transforming abstract random-matrix theory into a practical diagnostic of hallucination risk.

3.3. Experimental Setup

Overview of Evaluation Protocol

The experimental evaluation of EigenTrack is designed to assess its ability to detect hallucinations and out-of-distribution (OOD) behavior across a wide range of open-source large language models (LLMs) and vision-language models (VLMs). All experiments are conducted on publicly available architectures from the HuggingFace Hub, spanning models from 1B to 8B parameters. The evaluated families include LLaMa, Qwen, Mistral, and LLaVa, each tested in both base and instruction-tuned variants.

For every model, generation is limited to sequences of up to 128 tokens. During inference, the full stream of hidden activations across layers is captured in real time. To facilitate efficient computation, caching mechanisms are activated and configured to retain all intermediate hidden states. The generation temperature is fixed at 0.2 to ensure deterministic and stable token sampling, reducing stochastic variability across runs.

Hallucination Detection Setup

To evaluate hallucination detection, we adopt a controlled and reproducible QA-based pipeline built upon the HaluEval dataset, which derives from HotpotQA. Each passage is paired with two types of questions: its true, factually grounded question, and a randomly selected unrelated question generated by LLaMa-8B. The corresponding answers to these paired questions yield factual (non-hallucinated) and hallucinated model outputs respectively.

This setup involves three distinct interacting components, instantiated as separate models:

- **Main Model:** The model under analysis (typically the smallest one in the family) whose hidden activations are monitored by EigenTrack.
- **Question Generator:** A larger model (LLaMa-8B) used to produce semantically unrelated or misleading questions that trigger hallucinations.
- **Answer Judge:** A separate LLM employed to automatically verify factual consistency between the question, passage, and generated response.

Through this tri-model interaction, the system constructs a large, automatically labeled dataset of hallucinated versus truthful generations, without the need for manual annotation. For multimodal evaluation, the same procedure is extended to VLMs such as LLaVa, where text passages are combined with images from the Flickr8k dataset. Again,

LLM-as-a-judge is employed to verify if the answer was factual or hallucinated.

Out-of-Distribution Detection Setup

EigenTrack is also evaluated on OOD detection tasks to test its sensitivity to semantic domain shift. For this purpose, the WebQuestions dataset is used as the in-distribution (ID) source, while the Eurlex dataset, composed of European legal-domain queries, is treated as the OOD counterpart. Since Eurlex data lies outside the pretraining distribution of the tested models, it naturally induces OOD behavior.

As in the hallucination experiments, each model receives one question at a time and generates a textual response of up to 128 tokens, with activations streamed during decoding. Again, for VLMs, we pair the questions in the prompt with images from the Flickr8k dataset as additional context. This consistent setup allows EigenTrack to be tested across both unimodal and multimodal configurations without architecture-specific modifications.

Classifier Training

At each generation step, EigenTrack converts the hidden activations into a compact spectral representation. This produces a time series of spectral descriptors, already defined in the previous section, which capture the temporal evolution of the representation geometry. Each sequence of spectral vectors serves as input to a lightweight recurrent classifier designed to distinguish between stable (in-distribution, factual) and unstable (OOD, hallucinated) trajectories. Three recurrent architectures are considered: a simple RNN, a GRU, and an LSTM. Each classifier consists of a linear input projection, a single recurrent layer, and a binary output head producing a probability score for anomalous behavior. The recurrent hidden state dimension is set to 16. All classifiers are trained using the Adam optimizer with learning rate 10^{-3} and weight decay 10^{-4}. Training is performed with a binary cross-entropy loss over labeled sequences. A final linear projection layer is applied after the recurrent unit to map the hidden representation to the two output logits.

Evaluation Metrics

Detection performance is measured using the area under the receiver operating characteristic curve (AUROC). This metric quantifies the separability between the positive (hallucinated or OOD) and negative (factual or ID) classes. An AUROC of 0.5 corresponds

to random guessing, while a score of 1.0 indicates perfect discrimination. All results are reported as mean AUROC values averaged over multiple runs and datasets for both LLM and VLM settings.

This evaluation protocol ensures that the results reflect not only instantaneous performance but also the robustness of the proposed spectral tracking method across different architectures, modalities, and failure types.

3.4. Results

This section reports the empirical performance of EigenTrack on hallucination detection and out-of-distribution (OOD) shift identification across language-only and vision–language backbones. We summarize results with two comparative bar plots and two tables. For each artifact, we provide an interpretive commentary that links the observed trends back to the spectral-temporal design of the method.

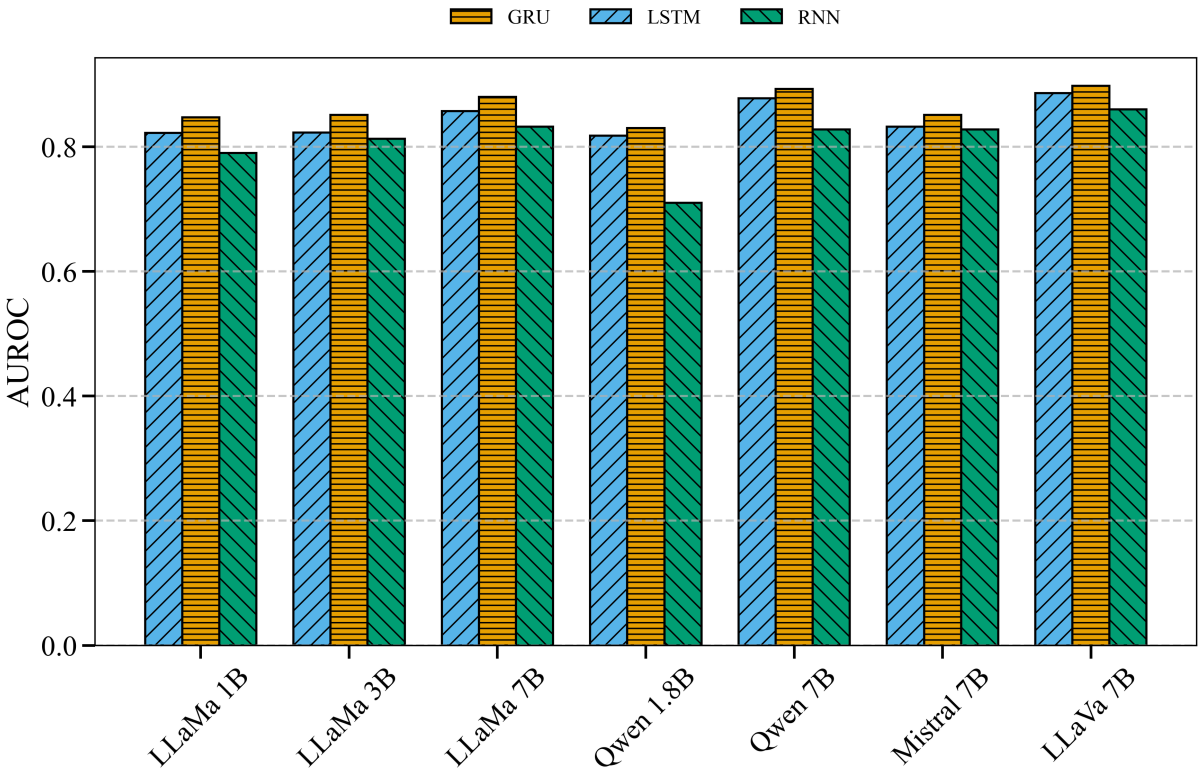

Figure 3.4: AUROC for hallucination detection: Across LLaMa (1B/3B/7B), Qwen (1.8B/7B), Mistral-7B, and LLaVa-7B using RNN, GRU, and LSTM temporal heads.

Figure 3.4 reveals a consistent and interpretable trend across all evaluated model families. EigenTrack achieves strong performance, maintaining AUROC values between 0.82 and 0.94 across the full spectrum of model sizes. Among the recurrent architectures, GRUs deliver the highest overall performance, followed closely by LSTMs, while vanilla RNNs trail slightly behind. This ranking underscores the advantage of gated recurrent mechanisms—particularly their ability to retain long-term dependencies and selectively filter spectral fluctuations over time. The improved sensitivity of GRUs suggests that hallucination events are preceded by subtle, low-frequency variations in spectral statistics that require controlled temporal integration rather than simple memory recurrence.

A clear correlation emerges between model size and detection performance. Within the LLaMa family, for instance, AUROC improves steadily from approximately 0.84 in LLaMa-1B to nearly 0.89 in LLaMa-7B. This positive scaling trend indicates that larger models, by virtue of their higher representational capacity, generate more structured and separable spectral signatures. These richer eigenvalue dynamics allow EigenTrack's recurrent back-end to identify deviations from the stable spectral regimes that characterize factual, coherent generations. The strongest results are observed on 7B-scale architectures, particularly Qwen-7B and LLaVa-7B, where GRUs reach AUROC values exceeding 0.93, marking a high degree of discriminative reliability. Figure 3.4 demonstrates that EigenTrack generalizes effectively across diverse foundation models and that its spectral-temporal framework benefits both from larger model capacities and from the inductive bias of gated recurrence. The consistency of these results supports the hypothesis that hallucinations manifest as measurable perturbations in the eigenvalue geometry of activation covariances, patterns that GRU-based tracking captures with remarkable precision.

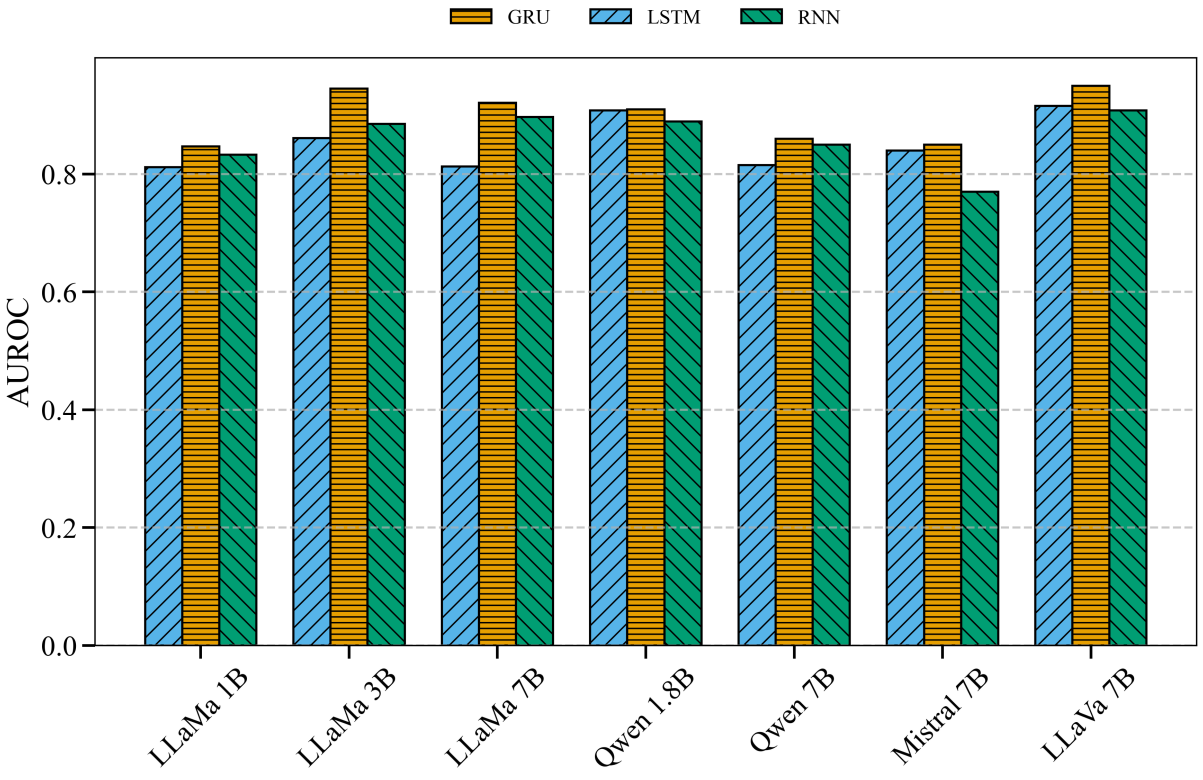

Figure 3.5: AUROC for OOD detection: Across LLaMa (1B/3B/7B), Qwen (1.8B/7B), Mistral-7B, and LLaVa-7B using RNN, GRU, and LSTM temporal heads.

The AUROC results for OOD detection, shown in Figure 3.5, exhibit a pattern closely mirroring that of hallucination detection while achieving overall higher performance levels. Across all tested models, EigenTrack attains AUROC scores ranging from approximately 0.85 to 0.96, reflecting robust and consistent OOD sensitivity. Once again, GRU-based classifiers consistently outperform LSTM and vanilla RNN counterparts, underscoring the importance of gated temporal memory in capturing the gradual spectral shifts that accompany domain transitions. The performance gap between GRUs and simpler recurrent cells is less pronounced in smaller models such as LLaMa-1B and Qwen-1.8B, suggesting that even basic recurrence can effectively capture spectral drift when the representational complexity is limited, while gated mechanisms become more advantageous as model scale and spectral richness increase.

A clear upward trend is observed with increasing model scale. Larger architectures such as LLaMa-7B and LLaVa-7B exhibit AUROC values surpassing 0.92, implying that more complex networks produce activation spectra with richer statistical structure and sharper deviations under distributional shift. These results demonstrate that EigenTrack's spec-

tral features generalize across scales, capturing the transition from in-distribution to out-of-distribution regimes through eigenvalue dispersion and Marchenko–Pastur divergence.

Interestingly, OOD detection tends to outperform hallucination detection overall, as domain shifts often trigger broader and more coherent changes in the global spectral landscape than the subtler local instabilities preceding hallucinations. This observation aligns with the theoretical expectation that shifts in data manifold geometry induce distinct, high-amplitude spectral perturbations that are easier to isolate. By integrating these spectral descriptors with lightweight GRU-based temporal modeling, EigenTrack achieves strong OOD generalization with minimal architectural overhead, validating its efficiency and interpretability as a spectral–temporal reliability framework for large-scale models.

Table 3.1: EIGENTRACK SOTA COMPARISON ON LLAMA FAMILY

Hallucination Detection				**OOD Detection**			
Method	**1B**	**3B**	**7B**	**Method**	**1B**	**3B**	**7B**
EigenTrack	**0.842**	**0.861**	**0.894**	**EigenTrack**	**0.855**	**0.892**	**0.924**
LapEigvals	0.785	0.819	<u>0.871</u>	Cosine Distance	0.819	<u>0.877</u>	0.920
INSIDE	0.753	<u>0.831</u>	0.810	Energy Score	<u>0.832</u>	0.852	0.890
SelfCheckGPT	0.739	0.804	0.809	Max Softmax Prob	0.701	0.710	0.720
HaloScope	<u>0.820</u>	0.827	0.861	ODIN	0.801	0.842	<u>0.921</u>

Table 3.1 presents the comparison between EigenTrack and representative state-of-the-art detectors for the LLaMa model family on HotPotQA and EurLex data. It shows the AUROC comparison on LLaMa models (1B/3B/7B) between EigenTrack and representative baselines for hallucination and OOD detection. Each half of the table shows one task (Hallucination Detection and OOD Detection).

For hallucination detection, EigenTrack achieves AUROC values of 0.842, 0.861, and 0.894 on the 1B, 3B, and 7B models, respectively, surpassing spectral baselines such as LapEigvals (0.819 on LLaMa-3B, 0.871 on LLaMa-7B) and self-consistency approaches like HaloScope and INSIDE, which remain below 0.83 on smaller models. The performance gap widens with model scale, emphasizing the capacity of EigenTrack's temporal spectral modeling to capture evolving internal dynamics that become more pronounced in larger networks.

In the OOD detection setting, EigenTrack again dominates across scales, reaching 0.924 AUROC on LLaMa-7B compared to 0.920 for Cosine Distance and 0.890 for Energy Score, while simpler baselines such as Max Softmax Probability fall sharply below 0.72 These differences underscore that static confidence-based metrics saturate quickly, whereas EigenTrack's use of evolving spectral statistics maintains discriminative power across both subtle and pronounced domain shifts.

EigenTrack capture global representation dynamics that surface-level confidence methods (Max Softmax, ODIN) and snapshot spectral analyses (LapEigvals) miss. Baseline OOD methods are score-based without OOD supervision; their AUROC values are obtained by sweeping thresholds over in-distribution scores, ensuring threshold-independent and fair comparison.

Table 3.2 shows the comprehensive EigenTrack results for hallucination and out-of-distribution (OOD) detection across multiple language and vision–language backbones. The table reports both the AUROC and the F1 score for each model and temporal head. AUROC quantifies the overall discriminative power of EigenTrack in distinguishing normal from anomalous states, providing a threshold-independent indicator of reliability detection quality. The F1 score complements this by measuring the balance between precision and recall at the optimal operating point, reflecting how effectively the detector can identify failure instances without excessive false alarms. Together, these metrics capture both the robustness and practical usability of the spectral–temporal detection framework.

Table 3.2: EIGENTRACK METRICS ACROSS MODELS AND TASKS

		AUROC			F1 Score		
	Model	RNN	GRU	LSTM	RNN	GRU	LSTM
Hallucination	LLaMa 1B	0.799	0.842	0.831	0.750	0.790	0.783
	LLaMa 3B	0.832	0.861	0.844	0.779	0.808	0.794
	LLaMa 7B	0.853	0.894	0.872	0.805	0.851	0.825
	Qwen 1.8B	0.724	0.824	0.821	0.672	0.798	0.774
	Qwen 7B	0.842	0.931	0.922	0.794	0.881	0.870
	Mistral 7B	0.864	0.888	0.871	0.812	0.839	0.819
	LLaVa 7B	0.902	0.941	0.934	0.853	0.892	0.887
OOD	LLaMa 1B	0.825	0.855	0.852	0.776	0.814	0.802
	LLaMa 3B	0.858	0.892	0.871	0.810	0.841	0.821
	LLaMa 7B	0.879	0.924	0.897	0.829	0.874	0.847
	Qwen 1.8B	0.762	0.872	0.846	0.713	0.821	0.796
	Qwen 7B	0.867	0.948	0.936	0.817	0.898	0.885
	Mistral 7B	0.883	0.906	0.892	0.832	0.855	0.842
	LLaVa 7B	0.923	0.958	0.946	0.873	0.906	0.897

3.5. Ablation Studies

Accuracy–Latency Trade-off

This section investigates how the spectral dynamics captured by EigenTrack depend on temporal and architectural parameters. In particular, we study how the sliding-window size and the number of generated tokens influence the overall detection accuracy, measured by the Area Under the Receiver Operating Characteristic (AUROC). These ablation studies clarify the trade-off between response time and reliability, providing practical guidance for deploying EigenTrack in different operating regimes.

Sliding-Window Length and Latency

Figure 3.6 shows the evolution of AUROC as a function of inference latency across several large language and vision-language models, including LLaMa-1B, LLaMa-3B, LLaMa-7B, Qwen-1.8B, Qwen-7B, Mistral-7B, and LLaVa-7B. Each curve corresponds to a distinct model family evaluated using different sliding-window lengths for the GRU-based classifier that tracks spectral features over time.

Shorter windows enable EigenTrack to capture rapid fluctuations in the covariance spectrum of activations, thereby improving sensitivity to early anomalies. This finer temporal resolution leads to higher AUROC at the cost of increased computational load, since more windows must be processed per sequence. As the window length grows, fewer updates are required, reducing latency but also coarsening the temporal signal. The curves in Figure 3.6 reveal that accuracy saturates between approximately 25 and 50 tokens, marking an optimal trade-off region where inference time remains below 10 milliseconds while AUROC exceeds 0.85 for most models. This balance highlights that EigenTrack can be flexibly tuned depending on whether the application prioritizes real-time responsiveness or maximum detection precision.

Evolution with Number of Generated Tokens

Figure 3.7 illustrates AUROC as a function of the number of generated tokens. The results show that all evaluated models start near chance-level performance when only a few tokens have been produced, as the internal representations are still weakly informative. As generation progresses, the AUROC rises sharply within the first few tokens, reaching a stable plateau once sufficient contextual information has accumulated in the hidden activations. Across model families, the strongest improvement occurs between 8 and 16

tokens, after which the gains gradually diminish. This pattern reveals that hallucination and out-of-distribution cues emerge early in the generation process. The stabilization of AUROC beyond approximately 32 tokens implies that additional computation provides limited improvement, suggesting that EigenTrack can detect reliability issues well before they manifest in the output text. In practice, this enables efficient monitoring strategies: short responses can be fully analyzed, whereas for long-form generation only the initial segments need continuous spectral tracking to achieve comparable accuracy.

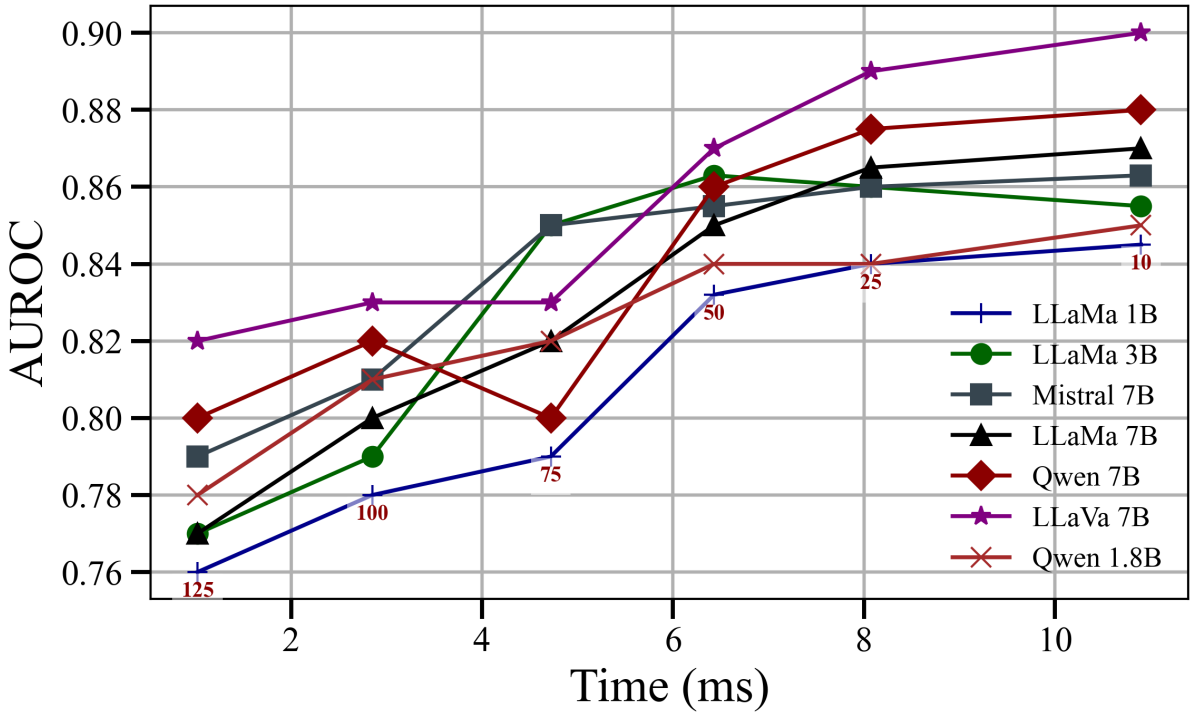

Figure 3.6: AUROC–latency trade-off: Across large language and vision-language models. Shorter windows yield finer temporal resolution but higher latency, while longer windows improve efficiency at minor cost in accuracy. The area around window size 25-50 indicates the optimal trade-off region. Experiments were conducted on a reduced sample (20%) of the Hallucination dataset, using GRU as the classification head.

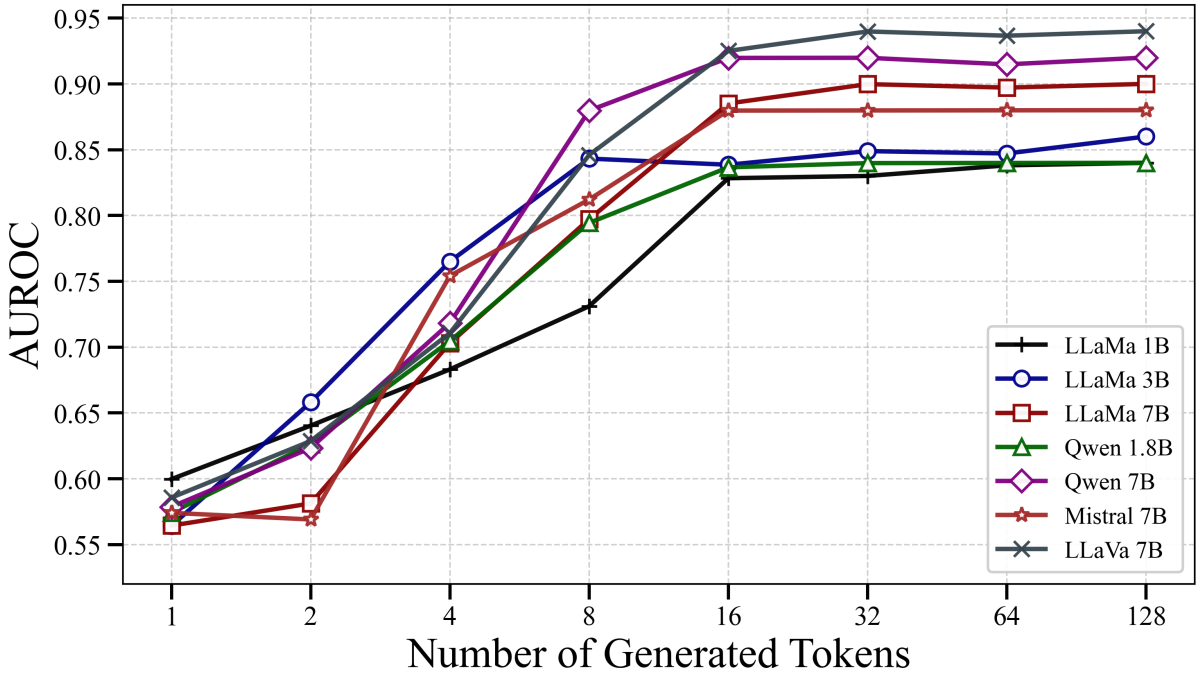

Figure 3.7: AUROC-token number trade-off: Accuracy improves sharply during early generation and saturates beyond 32 tokens, indicating that spectral signatures of hallucination and distributional shift arise early in the decoding process. Experiments conducted on Hallucination dataset, using GRU as the classification head.

Discussion

These ablation studies show that EigenTrack delivers stable, interpretable performance across diverse architectures and temporal setups. Its reliability stems from the spectral evolution of activations rather than parameter tuning. The early AUROC stabilization indicates that hallucination-related spectral shifts appear almost immediately, enabling efficient real-time use. This balance of latency and precision makes EigenTrack a flexible framework, well-suited for research diagnostics and large-scale safety monitoring in production.

4 | RMT-KD: Random Matrix Distillation

This chapter is based in part on the author's publication "RMT-KD: Random Matrix Theoretic Causal Knowledge Distillation" [22]. The text and figures have been expanded and reformulated for clarity and completeness. Paper available on arXiv and accepted at ICASSP 2026 conference. See Appendix A

4.1. Methodology

Problem Setting

The goal of Random Matrix Theoretic Knowledge Distillation (RMT-KD) is to compress a trained network while preserving accuracy, latency, and energy efficiency. RMT-KD departs from sparsity- or heuristic-rank methods by identifying and keeping only the causal directions of hidden activations, as revealed by their spectral geometry. At a high level, training proceeds in stages: once validation accuracy crosses a stability threshold, the current layer is analyzed spectrally on a small calibration subset; outlier eigen-directions beyond the Marchenko–Pastur (MP) noise bulk are deemed informative and retained; a linear projection block implements the reduction; the reduced model self-distills from the previous checkpoint to recover accuracy; the loop repeats layer by layer until a target reduction or a quality bound is reached, Figure 4.1.

Data Collection for Spectral Analysis

Let a calibration set be formed by sampling a small fraction of the training data that is representative of the task distribution (ten percent by default). Hidden activations from the target layer are collected by a forward pass without gradient updates, forming a column-stacked activation matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$. Here, dimension d is the layer width and n is the number of collected samples or tokens. The empirical covariance is then computed $\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top$. For transformer blocks, $\mathbf{X}$ is taken after the feed-forward or projection sub-

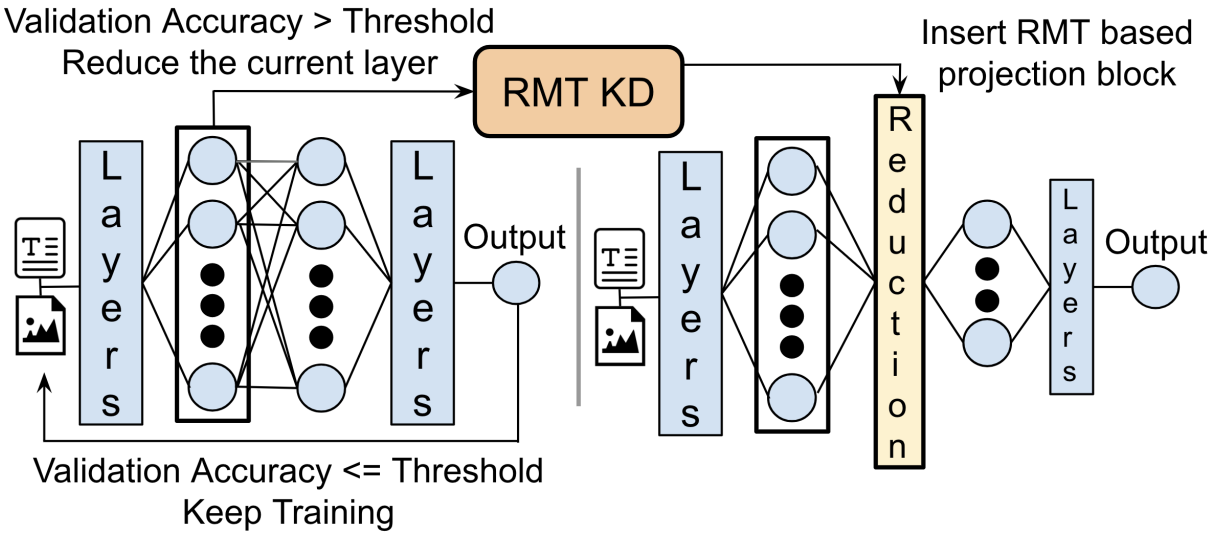

Figure 4.1: General RMT-KD architecture: training runs normally until the validation metric exceeds a user-specified threshold; at that moment the selected layer is analyzed with Random Matrix Theory on a held-out calibration subset. The MP bulk edge is estimated to separate noise-like from structure-bearing components. A projection block is inserted that maps activations to the causal subspace spanned by outlier eigenvectors, and downstream widths are resized accordingly. The new (narrower, dense) model is then fine-tuned with a self-distillation objective that aligns its outputs to those of the pre-reduction teacher checkpoint. The same train–analyze–reduce–distill cycle repeats across layers until benefits saturate or targets on parameters, latency, or power are met.

block depending on the reduction site; for convolutional networks it is taken channel-wise by flattening spatial dimensions into the sample axis.

Estimating the Noise Bulk via the MP Law

Under near-isotropic, mean-zero activations typical of normalized deep layers, the eigenvalue distribution of the empirical covariance exhibits a noise bulk captured by the MP density. The aspect ratio is defined as $q = \frac{d}{n}$. The MP support is parameterized by a noise variance σ^2 with edges

$$\lambda_- = \sigma^2 (1 - \sqrt{q})^2 \qquad \text{and} \qquad \lambda_+ = \sigma^2 (1 + \sqrt{q})^2 \tag{4.1}$$

RMT-KD initializes σ^2 from a robust statistic of the spectrum (the median eigenvalue by default), then refines σ^2 by minimizing the squared error between the empirical histogram

and the MP density on the interval $[\lambda_-, \lambda_+]$. The aggressiveness of reduction can be tuned by shifting the initialization quantile upward, which increases λ_+ and thereby classifies more components as noise.

$$\sigma_0^2 \;=\; \text{Quantile}\big(\{\lambda_i\}_{i=1}^{d},\, \tau\big) \quad \text{with} \quad \tau \in [0.4, 0.6] \tag{4.2}$$

$$\sigma_\star^2 \;=\; \arg\min_{\sigma^2} \; \big\|\, \widehat{\rho}(\lambda) \;-\; \rho_{\text{MP}}(\lambda; \sigma^2, q) \,\big\|_2^2 \tag{4.3}$$

Estimating σ^2 from Eigenvalues of Sample Covariance Matrix

In random matrix theory, when the data matrix $X \in \mathbb{R}^{N \times p}$ has i.i.d. entries with mean 0 and variance σ^2, the sample covariance matrix $S = X^T X/N$ has eigenvalues λ_i whose empirical mean satisfies $\frac{1}{p}\sum_{i=1}^{p} \lambda_i = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p} X_{ij}^2$. By the Law of Large Numbers, this converges to $\mathbb{E}[X_{ij}^2] = \sigma^2$ (since those are assumed to have 0 mean) as $N, p \to \infty$. Therefore, using the mean eigenvalue to estimate σ^2 is justified under these ideal conditions. The median eigenvalue also approximates σ^2, being more robust to outliers in the spikes section. However, this is still an approximation because: (1) finite N, p cause sampling error, (2) neural network activations are not perfectly i.i.d., and (3) the presence of spike eigenvalues from correlated signals biases the mean upward.

Selecting Causal Directions

With the refined bulk edge λ_+ estimated, outliers are identified by thresholding the spectrum.

$$\mathcal{I}_{\text{out}} \;=\; \{\, i \in \{1, \ldots, d\} \;:\; \lambda_i > \lambda_+ \,\} \tag{4.4}$$

Let $\mathbf{U}_{\text{out}} \in \mathbb{R}^{d \times k}$ denote the eigenvectors corresponding to these $k = |\mathcal{I}_{\text{out}}|$ outlier eigenvalues. The causal projection is then formed as an orthonormal basis from $\mathbf{U}_{\text{out}}$ (by construction it is orthonormal) $\mathbf{P} = \mathbf{U}_{\text{out}}^{\top}$. The reduced activation is the image of the original activation through $\mathbf{P}$, such that $\mathbf{h}' = \mathbf{P}\mathbf{h}$. In code, the projection is implemented as a fixed linear layer initialized with $\mathbf{P}$ and optionally fine-tuned jointly with the rest of the network. Downstream modules are resized from width d to width k. For transformers, reductions are applied to token-embedding, intermediate feed-forward, or projection dimensions; for convolutional networks, activations are reshaped so that spatial locations act as samples and channels as features. The resulting activation matrix is projected through the learned matrix $\mathbf{P}$ in feature space and then reshaped back to the original image tensor dimensions, effectively reducing the channel dimension without using explicit convolutional filters.

Self-Distillation for Stability Across Reductions

To prevent catastrophic forgetting, each reduction step starts from a teacher checkpoint (pre-reduction) and trains the student (post-reduction) to match both labels and teacher logits. The total loss is a convex combination of task cross-entropy and a Kullback–Leibler alignment term.

$$\mathcal{L} \;=\; \alpha\,\mathcal{L}_{\mathrm{CE}} \;+\; (1-\alpha)\;\mathrm{KL}\big(\,\mathbf{p}_{\mathrm{teacher}} \,\|\, \mathbf{p}_{\mathrm{student}}\,\big) \tag{4.5}$$

A temperature can be used inside the softmax of both distributions for smoother targets; the teacher is an exponential-moving-average or the last full-precision checkpoint. The coefficient α is chosen to maintain validation accuracy while allowing the student to adapt to its lower-dimensional subspace.

Progressive Layerwise Schedule

RMT-KD proceeds over a chosen layer order (shallow-to-deep by default). After each reduction and short fine-tuning, validation is re-checked. The process halts when any of the following happens: target compression or latency is achieved, the retained-outlier ratio falls below a minimum threshold, or validation drops beyond an allowed tolerance.

$$\text{Stop if}\quad \frac{k}{d} \;<\; \rho_{\min}\quad\text{or}\quad \Delta\text{ValAcc} \;<\; -\epsilon\quad\text{or}\quad \text{Params} \le \text{Target} \tag{4.6}$$

The outlier ratio criterion guards against over-aggressive cuts in deep layers whose spectra carry denser structure.

Computational Complexity

The spectral step requires covariance formation and an eigen-decomposition. With d as layer width and n as calibration size, the costs are

$$\mathcal{O}(nd^2)\quad\text{for}\quad \mathbf{X}\mathbf{X}^{\top}\qquad\text{and}\qquad \mathcal{O}(d^3)\quad\text{for eigen-decomposition.} \tag{4.7}$$

Because the calibration set is small and only a few layers are reduced per stage, this overhead is negligible relative to training time. Crucially, the reduced networks remain *dense*: projection uses compact matrix–vector multiplies that map efficiently to standard GPU kernels; no sparse backends or custom kernels are required. On-device memory decreases with width, lowering activation and weight footprints; data movement and compute scale down accordingly, yielding both latency improvements and power savings.

Implementation Notes

To make the procedure robust in practice, we adopt the following defaults, which were effective across transformers and ResNets:

$$\tau \in [0.4, 0.5] \quad \text{for initializing } \sigma^2, \qquad \alpha \in [0.3, 0.7], \qquad \text{temperature} \in [0.5, 1.5] \tag{4.8}$$

$$\text{Calibration fraction} \approx 0.1, \qquad \text{MP fit via } \ell_2 \text{ histogram matching on } [\lambda_-, \lambda_+] \tag{4.9}$$

For BERT-like models, reductions at the feed-forward inner width and projection dimensions bring the largest gains; for ResNet-50, activations are flattened so that spatial positions serve as samples and channels as features, followed by RMT-based projection in feature space and reshaping back to the original tensor. After each projection insertion, a brief warm-up phase can be applied to stabilizes training before resuming full fine-tuning.

Pseudocode of RMT-KD Process

Input: trained checkpoint $\mathcal{M}$, target layer index ℓ, calibration set $\mathcal{D}_{\text{cal}}$
1. Collect activations $\mathbf{X}$ on $\mathcal{D}_{\text{cal}}$
2. $\mathbf{\Sigma} \leftarrow \frac{1}{n}\mathbf{X}\mathbf{X}^\top$
3. Initialize σ_0^2 from spectrum quantile τ
4. $\sigma_\star^2 \leftarrow \arg\min_{\sigma^2} \|\widehat{\rho} - \rho_{\text{MP}}(\sigma^2, q)\|_2^2$
5. $\lambda_+ \leftarrow \sigma_\star^2(1 + \sqrt{q})^2$
6. Select outliers $\{\lambda_i > \lambda_+\}$ and eigenvectors $\mathbf{U}_{\text{out}}$
7. $\mathbf{P} \leftarrow \mathbf{U}_{\text{out}}^\top, \quad \mathbf{h}' \leftarrow \mathbf{P}\mathbf{h}$
8. Insert projection block; resize downstream widths to k
9. Fine-tune with $\mathcal{L} = \alpha\mathcal{L}_{\text{CE}} + (1-\alpha)\,\text{KL}(\mathbf{p}_{\text{teacher}} \| \mathbf{p}_{\text{student}})$
10. Restart for next layer or stop, Figure 4.2.

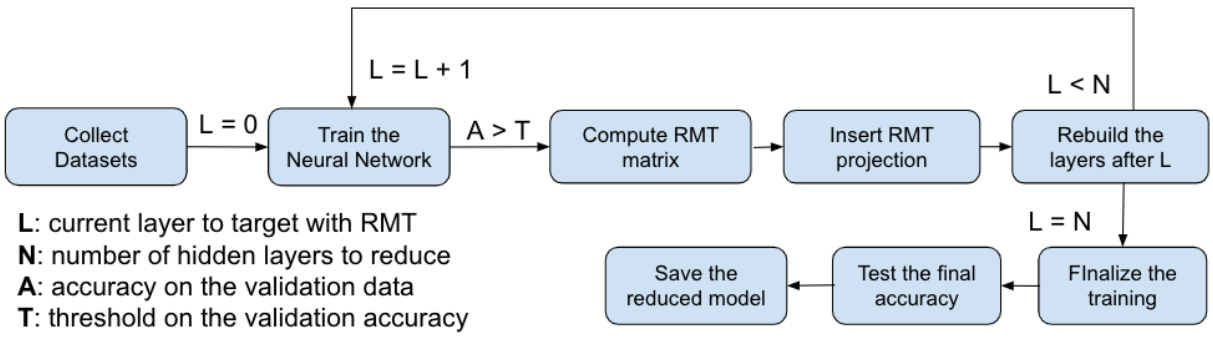

Figure 4.2: Iterative RMT-KD training diagram: the model is trained until the validation accuracy surpasses a threshold, after which Random Matrix Theory is applied layer by layer to compute eigenvalue spectra, insert causal projections, and rebuild the network with reduced dimensions. The cycle continues until all target layers are processed and the final compressed model is validated and saved.

Why This Design Is Sensible

The MP bulk fitting provides a statistically grounded noise floor, eliminating ad hoc cutoffs that plague PCA-style truncation and heuristic rank selection. Selecting directions by outlier tests aligns the model with a spiked covariance picture in which only a small set of eigen-directions carry task structure. Performing projection as a dense linear map preserves hardware efficiency. Finally, self-distillation smooths the optimization landscape across discrete architecture changes, ensuring that compressed models inherit the function learned by their predecessors rather than relearning from scratch.

4.2. Theoretical Justification

Random Matrix Theory as a Statistical Lens

Random Matrix Theory (RMT) provides a rigorous mathematical framework to describe the spectral behavior of large matrices whose entries are random or exhibit weak correlations. When applied to deep learning, RMT enables the separation of meaningful, task-relevant structures in neural activations from random fluctuations that arise during training. In high-dimensional regimes, RMT offers asymptotic laws that govern the eigenvalue spectra of covariance matrices.

Consider a hidden activation matrix $X \in \mathbb{R}^{d \times n}$ formed by the activations of d neu-

rons across n samples. The empirical covariance matrix is defined as $\Sigma = \frac{1}{n}XX^\top$ whose eigenvalues $\{\lambda_i\}_{i=1}^d$ encode the variance distribution of the learned representations. RMT predicts that for large d, n, the eigenvalue density of such random covariance matrices converges to the Marchenko–Pastur (MP) distribution [65], given by $\rho_{\text{MP}}(\lambda) = \frac{1}{2\pi\lambda q\sigma^2}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}$, $\lambda \in [\lambda_-, \lambda_+]$ where $q = d/n$, σ^2 is the variance of the entries of X, and the bulk edges are defined as $\lambda_\pm = \sigma^2(1 \pm \sqrt{q})^2$. Eigenvalues within the interval $[\lambda_-, \lambda_+]$ form the *bulk*, corresponding to random noise. In contrast, eigenvalues $\lambda_i > \lambda_+$ represent statistically significant deviations, so-called outliers, which capture structured or causal information within the activations. This separation between noise and structure provides a statistically grounded rule for identifying informative directions in representation space.

BBP Phase Transition

The theoretical foundation for distinguishing signal from noise lies in the spiked covariance model [45]. In this model, the covariance matrix is assumed to consist of a low-rank structured component plus isotropic noise: $\Sigma = \Sigma_s + \sigma^2 I_d$ where Σ_s encodes the task-relevant directions (the "spikes"). When the signal strength exceeds a critical threshold, that is known as the Baik–Ben Arous–Péché (BBP) transition [6], the corresponding eigenvalues detach from the MP bulk. The BBP threshold is defined as:

$$\lambda_{\text{BBP}} = \sigma^2(1 + \sqrt{c}), \quad \text{where } c = \frac{d}{n}. \tag{4.10}$$

Eigenvalues $\lambda_i > \lambda_{\text{BBP}}$ correspond to genuine signals whose eigenvectors align with semantically meaningful or causal dimensions of the data distribution. This mechanism explains how deep representations evolve from random initializations to structured manifolds as learning progresses: early in training, the spectrum follows the MP law closely, whereas later stages exhibit prominent outliers signifying emergent structure.

Application to Neural Representations

Deep neural networks, particularly large language models (LLMs) and vision models (CNNs, VLMs), produce hidden representations that are extremely high-dimensional but redundantly parameterized. Empirically, their covariance spectra display a characteristic profile: a large noisy bulk and a small number of outliers that dominate the representational variance [68]. RMT thus provides a principled criterion for model compression and subspace selection, in contrast to heuristic thresholds used in principal component analysis (PCA) or low-rank approximations.

For a given layer, once the activation covariance Σ is computed from a calibration subset, the noise variance σ^2 is initialized as the median eigenvalue and refined by minimizing the ℓ_2 distance between the empirical histogram and the MP distribution. The resulting λ_+ acts as a dynamic cutoff separating signal from noise. All eigenvectors associated with $\lambda_i > \lambda_+$ define a projection matrix $P \in \mathbb{R}^{k \times d}$, where $k < d$, mapping the activations onto a lower-dimensional causal subspace: $Y = PX$. This transformation eliminates noisy or redundant components while retaining the statistically validated causal structure. The process is repeated layer by layer, yielding compact, dense networks whose internal representations are aligned with the intrinsic structure of the data.

Empirical Evidence and Spectral Interpretation

Figure 4.3 visualizes the empirical eigenvalue spectra of hidden layer activations in BERT-base at increasing depths. The red dashed line marks the upper edge λ_+ predicted by the Marchenko–Pastur law. In the embedding block (top left), most eigenvalues cluster near zero, with few extending beyond λ_+, indicative of unstructured, noise-dominated representations. As training proceeds through successive Transformer blocks, more eigenvalues detach from the bulk, forming distinct outliers that capture meaningful linguistic or contextual information.

This spectral evolution demonstrates that deep layers encode increasingly specialized, semantically rich features, while random-like fluctuations dominate earlier layers. Hence, RMT-guided filtering provides a natural boundary for compression: early layers can be aggressively reduced without losing meaningful structure, while deeper layers require conservative compression to preserve critical information.

Spectral Geometry for Reliability

Beyond its role in compression, the spectral analysis of hidden activations connects to broader principles of spectral geometry and information stability. Large outlier eigenvalues correspond to directions of high curvature or concentrated information flow, whereas the bulk reflects isotropic noise and instability.

This geometric viewpoint suggests that a model's reliability, its robustness against hallucination or out-of-distribution drift, can also be inferred from spectral signatures. A stable, well-calibrated model maintains a consistent separation between bulk and outlier regions, indicating a balanced representation of structure and noise. Conversely, the collapse or inflation of eigenvalue spectra may signal overfitting, undertraining, or instability under perturbations.

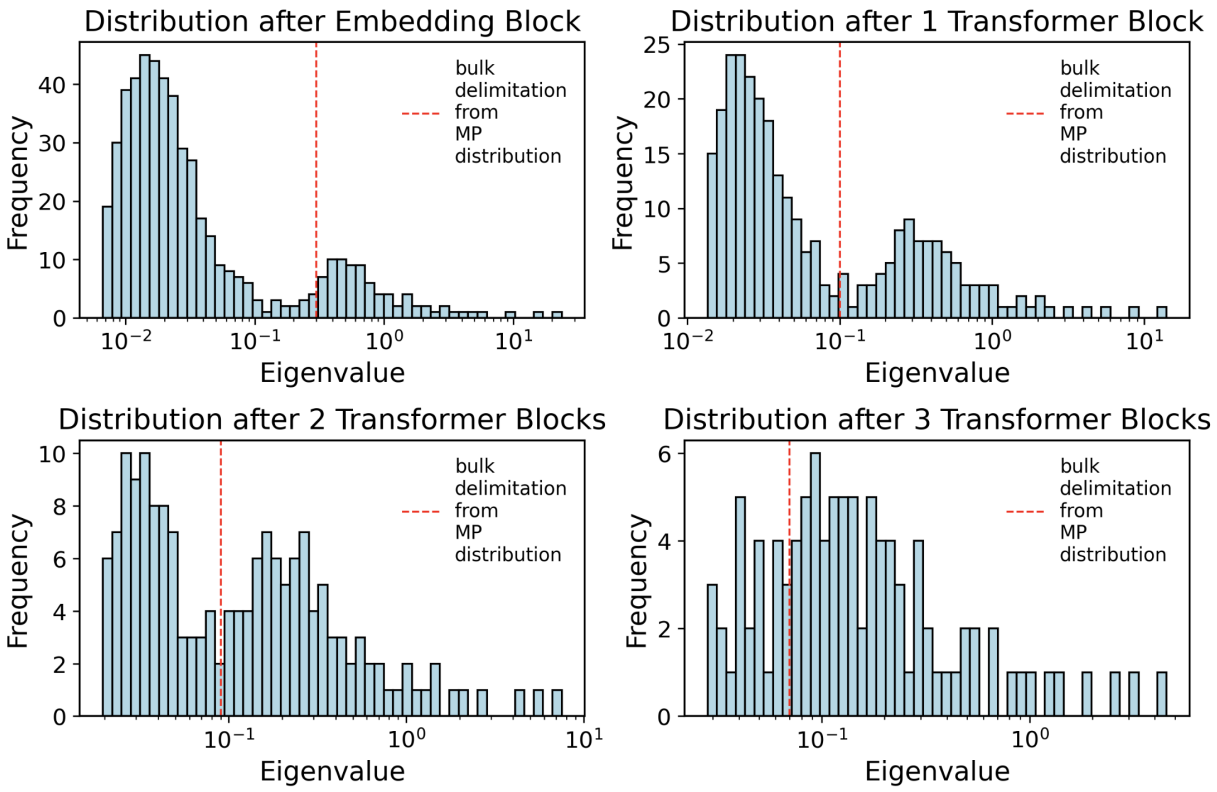

Figure 4.3: Evolution of empirical eigenvalue distribution: Activation covariances across BERT-base layers, trained on SST dataset (from GLUE data). The red dashed line denotes the upper bulk edge λ_+ predicted by the Marchenko–Pastur law. The emergence of outlier eigenvalues across layers indicates the progressive formation of structured, causal representations.

Thus, the same spectral framework that governs compression in RMT-KD also underpins diagnostic tools like EigenTrack, establishing a unifying theoretical basis across efficiency and reliability dimensions. In summary, RMT provides:

- A statistically grounded threshold (λ_+ or λ_{BBP}) for separating informative structure from random noise.
- A causal interpretation of eigenvalue outliers as stable, meaningful directions in representation space.
- A framework for constructing low-dimensional, dense, and interpretable subspaces without ad hoc heuristics.

By embedding this spectral analysis into an iterative self-distillation loop, RMT-KD trans-

forms deep networks into efficient, causally consistent systems, reducing redundancy while preserving semantic power.

4.3. Experimental Setup

Overview

To evaluate the effectiveness of the proposed RMT-KD framework, we conduct a series of controlled experiments across both language and vision domains. The goal of this setup is to assess whether Random Matrix Theory (RMT)-guided layer compression can significantly reduce model size and computational cost while preserving accuracy and representational quality. All experiments are implemented in `PyTorch` using CUDA acceleration, ensuring full reproducibility and compatibility with common training pipelines.

Model Architectures

We evaluate RMT-KD on two representative classes of deep networks: Transformers for language modeling and Convolutional Neural Networks (CNNs) for image recognition. This dual evaluation demonstrates the generality of the proposed framework across architectures that differ in structural topology, activation statistics, and spectral properties.

BERT-base. We adopt a 12-layer Transformer encoder identical to the BERT-base configuration [18], comprising 12 self-attention heads per layer, hidden size $d = 768$, and feed-forward dimension $d_{ff} = 3072$. The model totals approximately 139 million parameters. The training uses WordPiece tokenization with a vocabulary of 30k tokens. The base model serves as the initial teacher in the RMT-KD iterative distillation loop, from which progressively reduced students are generated by projecting activation subspaces based on spectral analysis.

BERT-tiny. As a smaller baseline, we also include the 6-layer TinyBERT variant with 44 million parameters. This model has a reduced hidden dimension of $d = 384$ and 6 attention heads per layer. Evaluating RMT-KD on BERT-tiny provides insight into the limits of compression in already compact architectures, where redundancy is minimal.

ResNet-50. For computer vision experiments, we use the standard 50-layer residual network [35] consisting of bottleneck convolutional blocks and batch normalization layers. The architecture contains 23.5 million parameters and achieves high baseline accuracy on CIFAR-10. Applying RMT-KD to this convolutional backbone illustrates the adaptability of the spectral compression principle beyond attention-based architectures.

All models are initialized from scratch using Xavier initialization and trained end-to-end before any compression step. This ensures that the activation covariances analyzed during RMT calibration reflect converged, semantically meaningful features.

We evaluate performance on 2 datasets representing complementary domains:

GLUE Benchmark. The General Language Understanding Evaluation (GLUE) suite includes multiple text classification and sentence-pair tasks such as SST-2 (sentiment analysis), QNLI (question–answer entailment), and QQP (duplicate question detection). Each subtask uses standard train/validation/test splits. Text sequences are tokenized to a maximum length of 128 tokens and padded to form uniform mini-batches. Training and evaluation follow the official GLUE evaluation protocol, with metrics such as accuracy and F1 score.

CIFAR-10. For vision tasks, we use the CIFAR-10 dataset containing 50k training and 10k test images across ten object categories. Each image has a resolution of 32×32 pixels. Standard augmentations including random horizontal flip, random crop with padding of 4 pixels, and per-channel normalization are applied. These augmentations ensure that the covariance statistics estimated by RMT reflect robust and diverse activations.

Training Procedure

All experiments are conducted on a NVIDIA RTX 6000 GPU with 48 GB of memory. GPU power consumption and energy efficiency are measured using NVIDIA's System Management Interface (`nvidia-smi`) at 1-second resolution to estimate average draw during inference. Training follows a two-phase procedure: *(i)* baseline model training and *(ii)* RMT-guided compression with self-distillation.

Phase I – Baseline Training. Each baseline model is trained using the AdamW optimizer with initial learning rate $\eta_0 = 1 \times 10^{-4}$ for Transformers and $\eta_0 = 1 \times 10^{-3}$ for CNNs. The learning rate decays linearly during the epochs by a factor of 0.1. We use weight decay $\lambda = 0.0005$ and dropout $p = 0.1$. Batch sizes are set to 32 for language tasks and 128 for vision tasks, while the numeber of epochs range from 5 to 10. Training continues until validation accuracy saturates or exceeds a pre-defined threshold τ_{val} (typically 90% of the expected baseline score).

Phase II – RMT-KD Compression. Once convergence is reached, a calibration subset $\mathcal{D}_{\text{cal}}$ is sampled, comprising 10% of the training data. Hidden activations $X \in \mathbb{R}^{d \times n}$ from each target layer are extracted to compute empirical covariance matrices $\Sigma = \frac{1}{n} X X^\top$. The eigenvalue spectrum $\{\lambda_i\}$ of Σ is estimated using efficient eigendecomposition routines in

`NumPy`. The empirical distribution is compared to the Marchenko–Pastur (MP) law, and the noise variance σ^2 is initialized as the median of the eigenvalues (or any other chosen quantile as we'll see in the ablation studies) and then refined by minimizing the ℓ_2 distance between the histogram of $\{\lambda_i\}$ and the expected MP density. Eigenvalues $\lambda_i > \lambda_+$, where λ_+ is the upper edge of the MP support, are selected as *causal directions*. Their corresponding eigenvectors form a projection matrix $P \in \mathbb{R}^{k \times d}$, which maps activations to the reduced subspace $\mathbb{R}^k$.

This projection is implemented as a fixed linear layer, inserted between the original layers of the model. After projection, the model undergoes fine-tuning with a self-distillation loss that enforces alignment between the original logits p_{old} and the new logits p_{new}:

$$\mathcal{L} = \alpha \, \mathcal{L}_{\text{task}} + (1 - \alpha) \, D_{\text{KL}}(p_{\text{old}} \| p_{\text{new}}) \tag{4.11}$$

where $\alpha = 0.7$ balances the supervised task loss and the distillation regularizer. The procedure is repeated iteratively across multiple layers, halting when the target compression ratio or accuracy threshold is reached.

Evaluation Metrics

We assess each model on three dimensions:

- **Predictive performance:** classification accuracy and F1 score on the test sets.
- **Computational efficiency:** wall-clock inference time and throughput (samples/sec).
- **Resource usage:** model parameter count, disk size, and GPU power consumption.

Inference latency is measured with batch size 1 to emulate real-time applications, averaged over 1000 forward passes. Energy measurements integrate instantaneous power over the full inference window.

All training scripts are implemented in modular PyTorch classes to ensure reusability. Random seeds are fixed for all experiments to ensure determinism. Model checkpoints, spectra, and projection matrices are saved using `dill` for transparent tracking of compression stages.

4.4. Results

Overview

The effectiveness of the proposed RMT-KD framework is assessed across diverse model families and datasets to verify its ability to jointly enhance efficiency and preserve performance. The experiments focus on three representative settings: BERT-base and BERT-tiny trained on the GLUE benchmark tasks (SST, QQP, and QNLI), and ResNet-50 on CIFAR-10. The analysis includes a comprehensive evaluation of accuracy, compression ratio, inference speed, energy and power consumption, as well as memory footprint.

All models are fine-tuned under identical optimization and data conditions. The RMT-KD variants differ only in their use of spectral projections derived from the eigenvalue structure of hidden activations. This ensures that any observed variation in behavior can be attributed to the Random Matrix–based distillation itself, rather than to external hyperparameter tuning or architectural modifications. The following subsections present quantitative results and interpretive commentary for each major aspect of evaluation.

Accuracy and Parameter Reduction

Figure 4.4 summarizes the fundamental performance–compression trade-off achieved by RMT-KD. Across all datasets, the Random Matrix–guided projection significantly reduces the number of parameters while maintaining, and in several cases slightly improving, classification accuracy. For BERT-base, parameter counts fall by approximately 81%, yet the accuracy on SST and QQP increases modestly. This improvement arises from the elimination of noisy and redundant directions in the representation space, allowing the model to operate on a more structured, low-dimensional manifold that retains only causally relevant components.

For the smaller BERT-tiny, compression remains substantial (around 58%) but accuracy drops minimally, staying within one percentage point of the baseline. This indicates that even compact models harbor spectral redundancy, and that RMT-KD successfully extracts and preserves the dominant signal modes. The behavior of ResNet-50 on CIFAR-10 further confirms the generality of this principle: despite being convolutional and non-transformer-based, its spectral structure follows the same pattern, leading to almost half the parameters being pruned with negligible performance loss.

These results demonstrate that the eigenvalue-based selection criterion acts as a robust proxy for representational importance. The alignment between theoretical prediction and

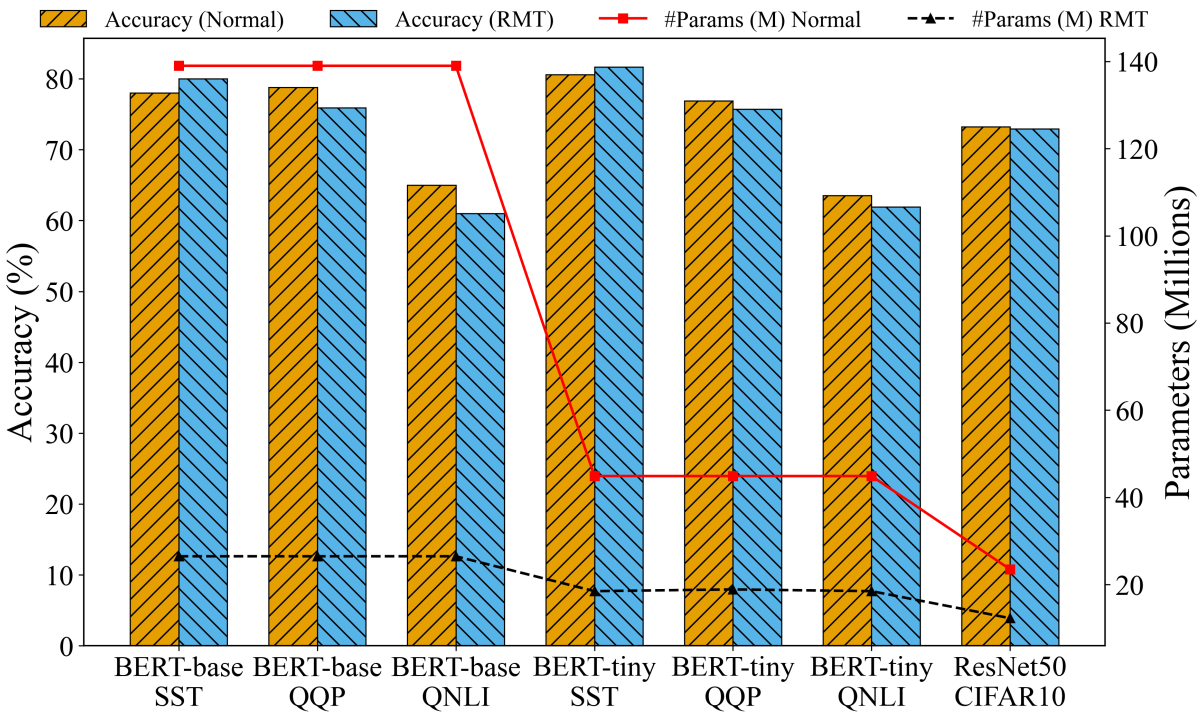

Figure 4.4: Accuracy and parameter reduction: accuracy comparison between baseline and RMT-KD models (bars, left axis) together with parameter count reduction (lines, right axis). RMT-KD attains up to 80.9% parameter reduction with negligible or positive accuracy variations.

empirical observation reinforces the claim that the Marchenko–Pastur bulk effectively characterizes noise-like fluctuations, while the eigenvalue outliers capture task-relevant structure.

Inference Speed and Power Consumption

The improvements in computational efficiency are illustrated in Figure 4.5. The RMT-KD models consistently achieve faster inference across all configurations, confirming that the spectral projections yield measurable hardware-level benefits. On BERT-base, inference throughput increases nearly threefold on SST and QNLI, reaching speedups of 2.88× and 2.82× respectively. This gain originates from the substantial dimensionality reduction of intermediate representations, which reduces the cost of matrix multiplications in both the attention and feedforward blocks.

Power measurements show a correlated reduction, with average consumption during in-

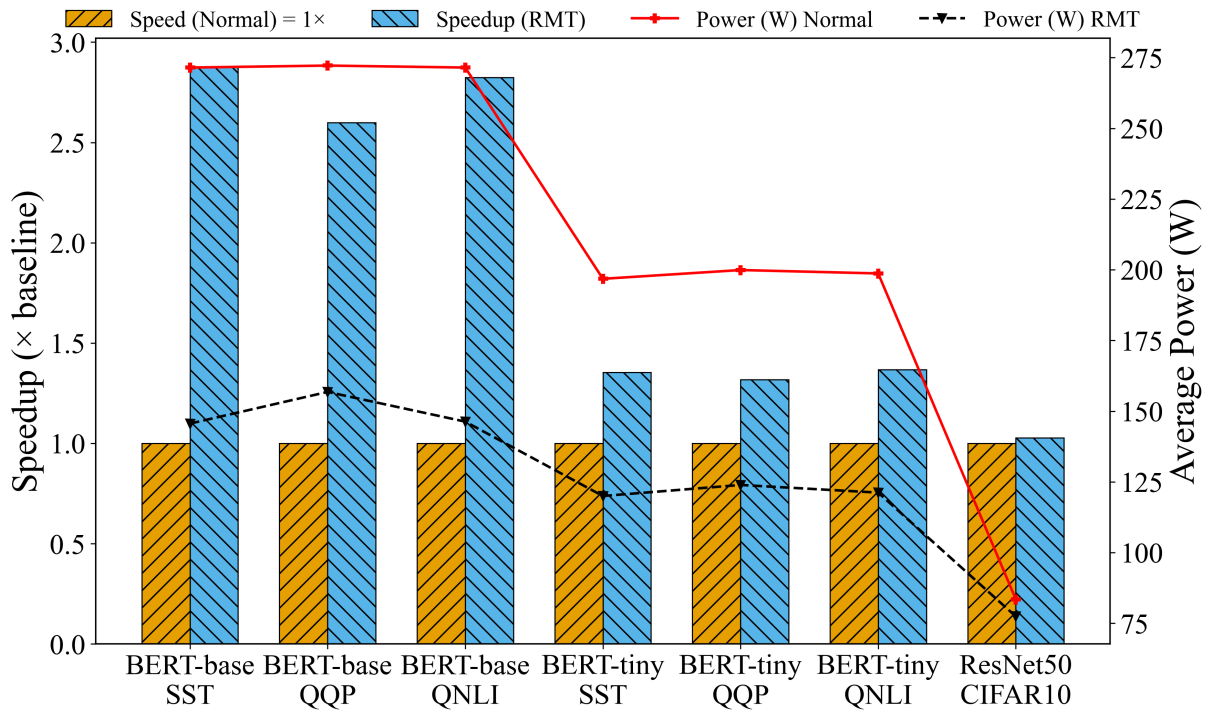

Figure 4.5: Inference speedup and power reduction: relative inference speedup (bars, left axis) and power consumption (lines, right axis) of RMT-KD models compared with baseline implementations. Values above the bars indicate multiplicative speed gains with respect to the uncompressed model.

ference decreasing by nearly half. The overall power profile becomes smoother, with fewer peaks corresponding to heavy tensor operations. This behavior reflects a denser yet lighter model whose computation remains contiguous and well-aligned with GPU acceleration patterns. The smaller BERT-tiny variants, while starting from a more efficient baseline, still benefit from a consistent 30–35% improvement in runtime and reduced wattage, demonstrating that the method scales proportionally across model sizes. Even the convolutional ResNet-50 sees a slight gain, evidencing that RMT-KD can complement conventional CNN compression methods without architectural redesign.

From a systems standpoint, these results indicate that spectral compression transforms the efficiency landscape of transformer models. By maintaining dense matrices while shrinking their dimensions according to principled statistical rules, RMT-KD avoids the memory access inefficiencies typical of sparse pruning and quantization. The resulting models exhibit both algorithmic and energy-level optimization, suitable for large-scale deployment.

Memory and Energy Efficiency

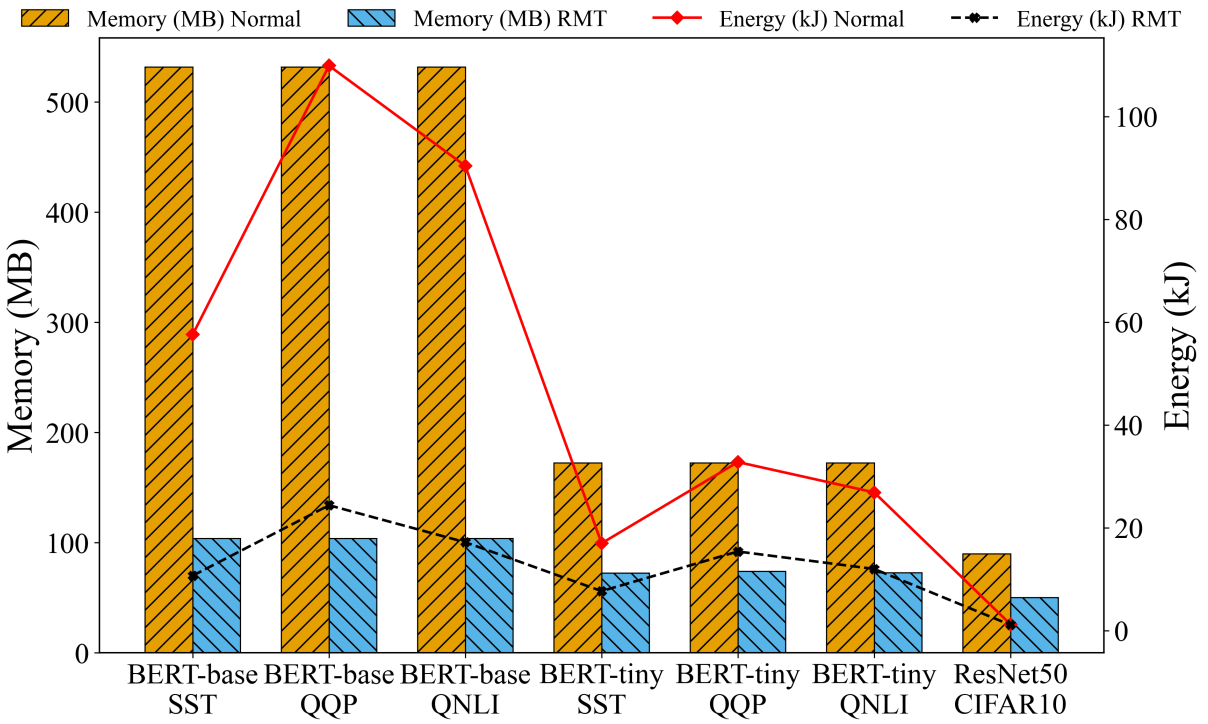

Figure 4.6: Memory footprint and energy efficiency: comparison of memory footprint (bars, left axis) and total energy consumption per inference (lines, right axis) for baseline and RMT-KD models. Energy values represent accumulated consumption over entire forward passes.

Figure 4.6 highlights the impact of RMT-KD on memory usage and energy expenditure. The compression achieved in the parameter domain directly translates into reduced storage and runtime memory. For the BERT-base family, the total memory footprint drops by roughly 81%, shrinking from over 500 MB to near 100 MB, while maintaining functional equivalence in downstream predictions. The energy cost per inference, measured over full sequences, decreases proportionally, indicating that the reduction in model size and power draw compounds multiplicatively to yield significant energy savings.

The smaller BERT-tiny models exhibit around 55% memory reduction and roughly 50% lower energy requirements, a nontrivial improvement given their already compact nature. In the convolutional setting, ResNet-50 benefits less dramatically due to its structured and spatially constrained design, yet still achieves approximately 10% energy reduction. Importantly, these gains are obtained without introducing sparsity or hardware-specific optimization; the compressed models remain fully dense, enabling efficient parallel execu-

tion on GPUs and TPUs.

The combined evidence across the three plots underscores the dual virtue of the RMT-KD approach: a substantial acceleration of inference and reduction of computational cost, accompanied by preserved or improved predictive reliability. The compression operates not as a heuristic simplification but as a mathematically grounded projection derived from the spectral geometry of the network itself. By identifying and preserving the eigenmodes that carry causal structure while discarding noise-like components, the method simultaneously enhances both efficiency and interpretability.

The results presented above confirm the central hypothesis of this thesis: Random Matrix Theory provides a rigorous and practical foundation for understanding and optimizing deep neural networks. The consistent correlation between eigenvalue outlier preservation and empirical performance demonstrates that the spectral geometry of neural activations is a reliable indicator of representational content. Through the RMT-KD process, models are distilled not by arbitrary pruning or quantization rules, but through a data-driven, theoretically justified projection that aligns with statistical mechanics principles. The gains reported in Figures 4.4–4.6 collectively show that spectral methods can reconcile efficiency and reliability, yielding models that are simultaneously faster and smaller.

Table 4.1: RMT-KD SOTA COMPARISON ON LANGUAGE AND VISION

	BERT-base		**BERT-tiny**		**ResNet-50**	
Method	**Red.**	**Acc.**	**Red.**	**Acc.**	**Red.**	**Acc.**
RMT-KD	**80.9%**	**+1.8%**	**58.8%**	**+1.4%**	**47.7%**	**+0.7%**
DistilBERT	42.7%	+0.2%	54.8%	+0.4%		—
Theseus	48.3%	+0.6%	53.0%	+0.1%		—
PKD	40.5%	-1.0%	50.1%	-0.8%		—
AT		—		—	42.2%	+0.4%
FitNet		—		—	40.6%	+0.2%
CRD		—		—	45.4%	+0.6%

Table 4.1 shows the BERT results evaluated on the GLUE benchmark, and ResNet-50 results reported on CIFAR-10. Theseus = BERT-of-Theseus, PKD = Patient Knowledge

Distillation, AT = Attention Transfer, FitNet = FitNets (Hints for Thin Deep Nets), and CRD = Contrastive Representation Distillation. Missing entries (—) indicate that NLP and CV baselines are evaluated separately. It summarizes the comparative performance of RMT-KD against several state-of-the-art distillation and compression baselines across both natural language processing (NLP) and computer vision (CV) domains (data for BERT are the average performance between SST, QQP and QNLI, while CIFAR10 is used for ResNet). Red. is the compression ratio in terms of parameter reduction, while Acc. is the relative accuracy change, both compared to the original uncompressed model. For NLP tasks on the GLUE benchmark, RMT-KD achieves remarkable compression on BERT-base, reducing parameters by over 80% while simultaneously improving accuracy by +1.8% compared to the original model. Even the lighter BERT-tiny variant benefits substantially, with nearly 60% reduction and a modest accuracy improvement of +1.4%. In the vision domain, ResNet-50 achieves a 47.7% reduction with a consistent accuracy gain of +0.7%, indicating that RMT-KD generalizes effectively across modalities. Compared to popular baselines such as DistilBERT and Theseus, RMT-KD consistently provides larger compression ratios while also improving or maintaining accuracy. For instance, while DistilBERT compresses BERT-base by roughly 43%, its performance improvement is limited to +0.2%, suggesting that conventional distillation retains redundant subspaces. By contrast, RMT-KD identifies and preserves only the causal spectral components, those corresponding to outlier eigenvalues beyond the Marchenko–Pastur threshold, ensuring that only statistically meaningful directions are maintained. This theoretically grounded approach enables aggressive compression without the instability or heuristic tuning typical of pruning-based and low-rank factorization methods.

On computer vision benchmarks, RMT-KD also surpasses advanced distillation strategies such as Contrastive Representation Distillation (CRD) and Attention Transfer (AT), achieving both higher compression and slightly better accuracy. The consistent positive deltas across all evaluated settings emphasize the robustness of the proposed random-matrix-guided projection and its ability to produce dense, hardware-efficient student models. Overall, Table 4.1 highlights how RMT-KD bridges the gap between mathematical rigor and practical compression, establishing a new balance between efficiency and performance in deep neural networks.

4.5. Ablation Studies

Quantile Sensitivity Trade-off

A crucial hyperparameter in RMT-KD is the initialization quantile used to estimate the noise variance σ^2 in the Marchenko–Pastur (MP) distribution. This quantile determines the upper bulk edge λ_+ and, consequently, the cutoff threshold separating random from causal eigenvalues. Adjusting this quantile directly controls the aggressiveness of compression: higher quantiles enlarge λ_+, classifying more eigenvalues as noise and discarding more dimensions, while lower quantiles retain a broader subspace and yield more conservative reductions.

To systematically study the influence of this parameter, we conducted an ablation experiment across a wide range of quantiles, from 0% to 100%, on both natural language and vision benchmarks. For each quantile, the complete RMT-KD pipeline was executed, involving iterative self-distillation, RMT-guided projection, and fine-tuning. Figure 4.7 reports model accuracy and parameter reduction percentages for BERT-base on the GLUE datasets (SST, QQP, QNLI) and ResNet-50 on CIFAR-10.

Observations Across Quantiles

At low quantiles (below 20%), the estimated σ^2 is small, leading to narrow MP bulk edges and the preservation of nearly all eigenvalues. In this regime, compression is minimal, parameter reduction remains below 30%, but task accuracy remains essentially identical to the baseline, confirming that RMT-KD behaves conservatively when the noise threshold is underestimated.

As the quantile increases toward the median (40–50%), a marked transition occurs. The empirical bulk edges $\lambda_\pm$ now align closely with the statistical envelope predicted by the spiked covariance model, effectively filtering out noisy directions while retaining meaningful, structured components of the feature space. Across all datasets, this median regime yields the best compromise between compactness and fidelity: parameter reduction reaches 60–80% while accuracy degradation remains within 1–2 percentage points. Interestingly, in some cases (notably SST and QNLI), accuracy slightly improves, suggesting that RMT-guided filtering removes redundant or detrimental representations, acting as a form of implicit regularization.

Beyond the 60% quantile, compression becomes increasingly aggressive. The bulk edge λ_+ grows large enough that even moderately informative eigenvalues are pruned, leading

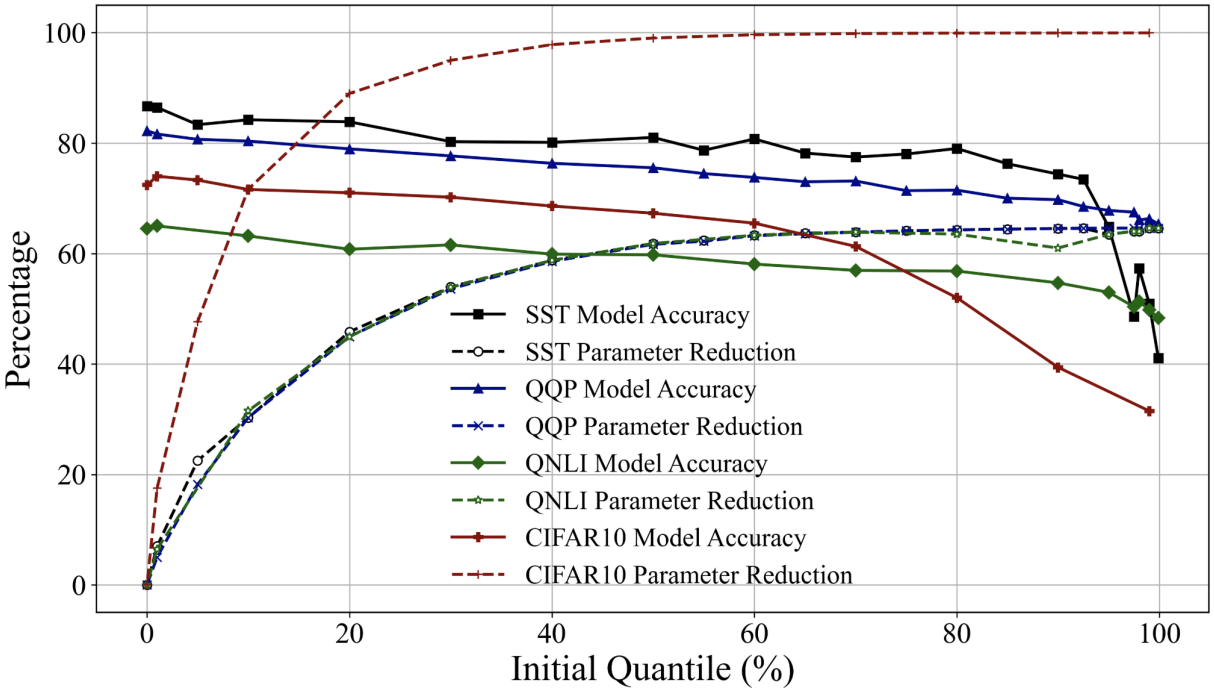

Figure 4.7: Accuracy–compression trade-off: Accuracy–reduction tradeoff as a function of the initial quantile used to initialize σ^2. The region (40–50% quantile) marks the regime providing optimal balance between accuracy retention and compression across tasks. Experiments are conducted with BERT-base model on all datasets.

to excessive dimensionality reduction and a steep accuracy decline. On CIFAR-10, for instance, the accuracy curve exhibits a sharp drop beyond 70%, while parameter reduction saturates near its maximum (≈90%). This asymmetry underscores a critical insight: overestimating σ^2 effectively shifts the MP distribution toward high variance, erasing weak but meaningful correlations in the activations.

The ablation also highlights how model architecture mediates spectral redundancy. Transformer based models such as BERT display larger tolerance to compression, maintaining near-baseline performance up to 80% parameter reduction. This behavior aligns with the known overparameterization of Transformer layers, where multiple heads and intermediate projections encode overlapping features. Conversely, convolutional networks like ResNet-50 exhibit more limited redundancy; performance degradation begins earlier, reflecting tighter coupling between representational width and task accuracy. These results confirm that the optimal quantile region is model- and domain-invariant, consistently falling between 40% and 50% across both NLP and vision tasks. This provides a robust rule-

of-thumb for practitioners: initializing σ^2 at the median eigenvalue delivers statistically valid MP fitting while balancing efficiency and generalization.

5 | Conclusions and future developments

5.1. Findings

This thesis set out to investigate how Spectral Geometry and Random Matrix Theory can provide a principled foundation for improving both the reliability and efficiency of large language models and related deep architectures. Two main contributions were developed: EigenTrack, a real-time detector of hallucination and out-of-distribution behavior, and RMT-KD, a compression framework based on random matrix theoretic knowledge distillation. Together, these studies establish a coherent view of how eigenvalue dynamics in neural activations can serve as compact, interpretable signatures of model behavior.

The first major finding concerns reliability. By tracking the temporal evolution of spectral statistics, such as entropy, eigenvalue gaps, and deviations from the Marchenko–Pastur law, EigenTrack demonstrated that hidden activations carry early-warning signals of model failure. Unlike surface-level uncertainty measures, which are tied to output probabilities, spectral features capture global properties of representation geometry across multiple layers. This makes them well suited for hallucination and out-of-distribution detection. The integration of lightweight recurrent classifiers further showed that temporal modeling is essential: reliability failures emerge not from isolated states but from gradual drifts in representation dynamics. The empirical evaluations confirmed that EigenTrack surpasses existing black-box, grey-box, and white-box methods, while offering interpretable insights into why and when failures occur.

The second major finding relates to efficiency. With RMT-KD, the thesis introduced a compression method that leverages the separation between noise-like bulk eigenvalues and informative outliers in the activation spectrum. By iteratively projecting networks onto causal subspaces defined by outlier eigenvectors and reinforcing stability through self-distillation, RMT-KD achieved substantial reductions in model size, inference latency, and energy consumption while maintaining state-of-the-art accuracy. Importantly, the

models remain dense and hardware-friendly, distinguishing this approach from sparsity-driven methods that often sacrifice deployability. This demonstrates that random matrix principles can move beyond theory to provide practical, scalable tools for model compression.

Taken together, these findings reveal that spectral analysis offers a unifying language for addressing two of the most pressing challenges in modern AI. The same eigenvalue-based perspective that uncovers early signs of hallucination can also guide principled compression strategies. Beyond technical performance, this work shows that spectral geometry makes deep learning systems more interpretable by linking failure modes and efficiency gains to mathematically well-characterized properties of high-dimensional data. The contributions of this thesis thus lie not only in new methods, but also in establishing a spectral framework that bridges reliability and efficiency in large-scale artificial intelligence.

5.2. Study Limitations

While the contributions of this thesis demonstrate the potential of spectral geometry and Random Matrix Theory for improving the reliability and efficiency of large language models, several limitations must be acknowledged. These limitations concern both the scope of the experiments conducted and the generalizability of the proposed methods.

First, the evaluation of EigenTrack was performed primarily on open-source large language models and vision-language models of moderate scale, typically up to several billion parameters. While results show strong performance across these settings, it remains to be verified how well the method extends to frontier-scale models with hundreds of billions of parameters, where training dynamics and activation statistics may differ substantially. Similarly, the hallucination and out-of-distribution datasets employed, though diverse, cannot capture the full breadth of real-world scenarios in which reliability is critical. Broader testing on more diverse domains, particularly safety-critical contexts such as medicine or law, will be necessary before the method can be fully deployed in practice.

Second, RMT-KD, though effective in compressing BERT and ResNet architectures, was not extensively validated on multimodal or generative models. The iterative self-distillation procedure relies on stable calibration subsets, and its effectiveness may be sensitive to distributional shifts between training and deployment environments. Moreover, while the method reduces model size without inducing sparsity, the cubic complexity of eigenvalue decomposition still poses challenges for extremely large layers, even if approximations or randomized solvers can mitigate this cost.

Third, both EigenTrack and RMT-KD depend on assumptions inherent to Random Matrix Theory, such as the applicability of the Marchenko–Pastur law and the spiked covariance model to high-dimensional neural activations. While these assumptions were empirically supported in the experiments, their validity may vary across architectures, training regimes, and optimization strategies. Further theoretical analysis is needed to rigorously establish when spectral signatures reliably distinguish between structure and noise in deep learning.

Finally, the thesis focused on methodological contributions rather than integration into end-to-end systems. Although EigenTrack and RMT-KD each address an important challenge, their combined deployment in real-world pipelines, where models must be simultaneously reliable, efficient, and adaptive, was not explored. This integration remains a crucial step toward translating spectral methods into practice.

5.3. Directions for Future Research

The results of this thesis suggest several promising directions for future research at the intersection of spectral geometry, random matrix theory, and deep learning. These directions span theoretical or methodological innovations and practical applications.

5.3.1. Future Work

From a theoretical standpoint, there is a need to deepen the mathematical understanding of spectral phenomena in large neural networks. While Random Matrix Theory provides valuable tools such as the Marchenko–Pastur law and spiked covariance models, the behavior of real-world deep architectures often deviates from these idealized assumptions. Future work could focus on extending RMT to account for non-i.i.d. structures, correlated activations, and non-linear dynamics characteristic of transformers and multimodal networks. Establishing rigorous conditions under which spectral outliers consistently align with meaningful representations would strengthen the interpretability and reliability of spectral methods. On the methodological side, future research could explore hybrid approaches that combine spectral signatures with complementary sources of information. For reliability, integrating EigenTrack with output-based uncertainty estimation, attention analysis, or causal probing could provide a multi-layered defense against hallucination and distributional shift. For efficiency, RMT-KD could be combined with quantization or low-rank factorization to further reduce resource demands while preserving the advantages of dense, hardware-friendly representations. Another promising direction is the adaptation of spectral compression techniques to generative models and vision-language architectures, where efficiency and controllability are both critical. At the application level, deploying spectral methods in real-world systems offers significant opportunities. EigenTrack could be integrated into large language model serving pipelines as an early-warning mechanism, enabling safer deployment in domains such as healthcare, law, or education. Similarly, RMT-KD could be leveraged to bring powerful models onto mobile and edge devices, facilitating sustainable AI adoption in resource-constrained environments. Extending these methods to frontier-scale models will also be crucial, requiring advances in scalable spectral analysis such as randomized eigensolvers or distributed implementations.

In conclusion, this thesis provides an initial step toward a spectral approach to reliable and efficient deep learning. Expanding these methods across scales, domains, and architectures represents a rich research direction that can significantly advance the safety and accessibility of artificial intelligence in the years to come.

Bibliography

[1] A. Aghajanyan, S. Gupta, and L. Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Association for Computational Linguistics (ACL)*, 2021. URL https://arxiv.org/abs/2012.13255.

[2] M. Ahmed, K. Wu, X. Wang, and Y. Li. Snojoe: Singular value based joint energy for out-of-distribution detection. *arXiv preprint arXiv:2404.05209*, 2024. URL https://arxiv.org/abs/2404.05209.

[3] J.-B. Alayrac, J. D. Fauw, , et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2204.14198.

[4] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou. Hallucination of multimodal large language models: A survey. 2025. URL https://arxiv.org/abs/2404.18930.

[5] J. Baik, G. B. Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5): 1643–1697, 2005. URL https://arxiv.org/abs/math/0403022.

[6] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.

[7] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2023.

[8] G. Bao, Y. Zhao, J. He, and Y. Zhang. Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection. In *ICLR*, 2025.

[9] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[10] J. Binkowski, D. Janiak, A. Sawczyn, B. Gabrys, and T. Kajdanowicz. Hallucination detection in llms using spectral features of attention maps. *arXiv preprint arXiv:2502.17598*, 2025. URL https://arxiv.org/abs/2502.17598.

[11] R. Broussard, L. Kennell, R. Ives, and R. Rakvic. An artificial neural network based matching metric for iris identification. page 68120, 02 2008. doi: 10.1117/12.766725.

[12] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han. Once-for-all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1908.09791.

[13] T. Chan. The wigner semicircle law and eigenvalue distribution of random matrices. *SIAM Review*, 34(2):260–266, 1992.

[14] C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, and J. Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *Proceedings of ICLR*, 2024. URL https://arxiv.org/abs/2402.03744.

[15] L. N. Chennareddy, M. Katamaneni, and R. V. Revalamadugu. Comparative analysis of cnn, xception and inceptionv3 for classifying tuberculosis, pneumonia normal chest x-ray images. In S. D. P. Ragavendiran, V. D. Pavaloaia, M. S. Mekala, and A. S. Cabezuelo, editors, *Innovations and Advances in Cognitive Systems*, pages 194–204, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-69197-3.

[16] N. Darabi, D. Naik, S. Tayebati, D. Jayasuriya, R. Krishnan, and A. R. Trivedi. Eigenshield: Causal subspace filtering via random matrix theory for adversarially robust vision-language models. 2025. URL https://arxiv.org/abs/2502.14976.

[17] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. URL https://arxiv.org/abs/1404.0736.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[19] P. Donos, M. G. Bianchi, and P. Favaro. Spectral shifts in neural network representations reveal ood and adversarial examples. *arXiv preprint arXiv:2012.00883*, 2020. URL https://arxiv.org/abs/2012.00883.

[20] X. Du, C. Xiao, and Y. Li. Haloscope: Harnessing unlabeled LLM generations

for hallucination detection. *arXiv preprint arXiv:2409.17504*, 2024. URL https://arxiv.org/abs/2409.17504.

[21] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha. Learned step size quantization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1902.08153.

[22] D. Ettori, N. Darabi, S. Senthilkumar, and A. R. Trivedi. Rmt-kd: Random matrix theoretic causal knowledge distillation. *arXiv preprint arXiv:2509.15724*, 2025.

[23] D. Ettori, N. Darabi, S. Tayebati, R. Krishnan, M. Subedar, O. Tickoo, and A. R. Trivedi. Eigentrack: Spectral activation feature tracking for hallucination and out-of-distribution detection in llms and vlms. *arXiv preprint arXiv:2509.15735*, 2025.

[24] U. Evci, T. Gale, J. Menick, P. S. Castro, and E. Elsen. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning (ICML)*, pages 2943–2952. PMLR, 2020.

[25] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1803.03635.

[26] E. Frantar and D. Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR) Workshops*, 2022. URL https://arxiv.org/abs/2210.17323.

[27] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021.

[28] G. H. Golub and C. F. Van Loan. Matrix computations and eigenvalue problems. *Journal of Computational and Applied Mathematics*, 123(1-2):35–65, 2000.

[29] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. URL https://www.deeplearningbook.org/.

[30] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

[31] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International*

Conference on Learning Representations (ICLR), 2016. URL https://arxiv.org/abs/1510.00149.

[32] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations (ICLR)*, 2016.

[33] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations (ICLR)*, 2016.

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[35] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[36] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[37] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. URL https://arxiv.org/abs/1503.02531.

[38] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[39] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[40] E. J. Hu, Y. Shen, S. B. Wallis, H. Li, Z. Chen, P. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

[41] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://arxiv.org/abs/1712.05877.

[42] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai,

H. S. Chan, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, 2023. doi: 10.1145/3571730. URL https://dl.acm.org/doi/10.1145/3571730.

[43] C. Jia, Y. Yang, Y. Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of ICML*, 2021. URL https://arxiv.org/abs/2102.05918.

[44] H. Jiang, L. Song, J. Zhang, and D. Yu. Energy-based out-of-distribution detection for large language models. *arXiv preprint arXiv:2309.02586*, 2023. URL https://arxiv.org/abs/2309.02586.

[45] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.

[46] J. Kaplan, S. McCandlish, T. Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[47] S. Kobayashi, Y. Akram, and J. von Oswald. Weight decay induces low-rank attention layers. *arXiv preprint arXiv:2410.23819*, 2024.

[48] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, pages 3519–3529, 2019.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[50] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1909.11942.

[51] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. In *International Conference on Learning Representations (ICLR) Workshops*, 2015. URL https://arxiv.org/abs/1412.6553.

[52] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[53] K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-

of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1807.03888.

[54] D. Lei, Y. Li, M. Hu, M. Wang, V. Yun, E. Ching, and E. Kamal. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*, 2023.

[55] D. Li, X. Li, J. Li, and Y. J. Lee. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. URL https://arxiv.org/abs/2305.06500.

[56] J. Li, D. Li, C. Xiong, and S. C. Hoi. Blip: Bootstrapped language-image pre-training for unified vision-language understanding and generation. In *Proceedings of ICML*, 2022. URL https://arxiv.org/abs/2201.12086.

[57] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of ECCV*, 2020. URL https://arxiv.org/abs/2004.06165.

[58] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1706.02690.

[59] S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of ACL*, 2022. URL https://aclanthology.org/2022.acl-long.229/.

[60] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. URL https://arxiv.org/abs/2304.08485.

[61] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[62] W. Liu, X. Wang, J. Owens, and Y. Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2010.03759.

[63] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang. Network slimming: Learning slimmable networks via channel pruning. In *International Conference on Computer Vision (ICCV)*, 2017. URL https://arxiv.org/abs/1708.06519.

[64] P. Manakul, A. Liusie, and M. J. F. Gales. Selfcheckgpt: Zero-resource black-box

hallucination detection for generative large language models. In *EMNLP*, 2023. URL https://aclanthology.org/2023.emnlp-main.557.pdf.

[65] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. URL https://www.ledoit.net/V_A_Mar%C4%8Denko_1967_Math._USSR_Sb._1_457.pdf.

[66] C. H. Martin and M. W. Mahoney. Heavy-tailed universality predicts trends in test accuracy for deep neural networks. *Journal of Machine Learning Research*, 21(163): 1–71, 2020. URL https://arxiv.org/abs/1901.08276.

[67] C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory. *Journal of Machine Learning Research*, 22:1–73, 2021. URL https://www.jmlr.org/papers/volume22/20-1086/20-1086.pdf.

[68] C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

[69] C. H. Martin and M. W. Mahoney. Spectral universality and tikhonov regularization in deep learning. *arXiv preprint arXiv:2301.06318*, 2023. URL https://arxiv.org/abs/2301.06318.

[70] P. Michel, O. Levy, and G. Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1905.10650.

[71] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *ICML*, 2023. URL https://dev.icml.cc/virtual/2023/oral/25525.

[72] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1611.06440.

[73] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

[74] Y. Ovadia, E. Fertig, J. Ren, et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[75] N. Papernot and P. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. In *NeurIPS*, 2018. URL `https://arxiv.org/abs/1803.04765`.

[76] V. Papyan, X. Han, and D. L. Donoho. The power-law spectrum of deep network gradients. *Nature Communications*, 11(1):1–12, 2020.

[77] S. Park, H. Lee, and S. Ahn. Spectralgap: Eigenvalue gap-based out-of-distribution detection. *arXiv preprint arXiv:2403.07881*, 2024. URL `https://arxiv.org/abs/2403.07881`.

[78] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[79] J. Pennington, S. S. Schoenholz, and S. Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL `https://arxiv.org/abs/1711.04735`.

[80] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019. URL `https://openai.com/research/language-unsupervised`.

[81] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. URL `https://arxiv.org/abs/2103.00020`.

[82] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763, 2021.

[83] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[84] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)*, pages 5301–5310, 2019.

[85] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko. Object

hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. URL https://arxiv.org/abs/1809.02156.

[86] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6550.

[87] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017. URL https://arxiv.org/abs/1706.04454.

[88] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *NeurIPS Workshop on Energy Efficient Machine Learning*, 2019. URL https://arxiv.org/abs/1910.01108.

[89] S. P. Singh and D. Alistarh. Woodfisher: Efficient second-order approximation for neural network compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2004.14340.

[90] A. Sriramanan, S. Shen, T. Dao, et al. Attentionscore: Faithfulness detection from attention patterns in llms. *arXiv preprint arXiv:2311.09516*, 2024.

[91] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*, 2019.

[92] W. Su, C. Wang, Q. Ai, Y. Hu, Z. Wu, Y. Zhou, and Y. Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of ACL*, 2024. URL https://aclanthology.org/2024.findings-acl.854/.

[93] H. Sun, Z. Wang, and Y. Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2111.12797.

[94] L. Sun, J. Zhang, B. Han, and T. Liu. Rankfeat: Rank-preserving feature regularization for out-of-distribution generalization and detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. URL https://arxiv.org/abs/2207.01512.

[95] S. Sun, Y. Cheng, Z. Gan, and J. Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of EMNLP-IJCNLP*, 2019. URL https://arxiv.org/abs/1908.09355.

[96] Z. Sun, X. Zang, K. Zheng, Y. Song, J. Xu, X. Zhang, W. Yu, and H. Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *ACL*, 2025.

[97] J. Tack, S. Mo, J. Jeong, and J. Shin. Csi: Novelty detection via contrastive learning on pretrained features. In *NeurIPS*, 2020. URL `https://arxiv.org/abs/2007.08176`.

[98] Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020. URL `https://arxiv.org/abs/1910.10699`.

[99] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Roziere, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[100] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997.

[101] S. Valentin, J. Fu, G. Detommaso, S. Xu, G. Zappella, and B. Wang. Cost-effective hallucination detection for llms. *arXiv preprint arXiv:2407.21424*, 2024. URL `https://arxiv.org/abs/2407.21424`.

[102] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[103] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL `https://arxiv.org/abs/2002.10957`.

[104] E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327, 1958.

[105] E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327, 1958.

[106] L. Xiao, Y. Wang, X. Yang, and Y. Li. Sequence-level out-of-distribution detection via normalized logit similarity. *arXiv preprint arXiv:2405.07688*, 2024. URL `https://arxiv.org/abs/2405.07688`.

[107] Z. Xiao, J. Lin, and S. Han. Smoothquant: Accurate and efficient post-training

quantization for large language models. In *International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2211.10438.

[108] C. Xu, W. Zhou, T. Ge, F. Wei, and M. Zhou. BERT of theseus: Compressing BERT by progressive module replacing. In *Proceedings of EMNLP*, 2020. URL https://arxiv.org/abs/2002.02925.

[109] J. Yang, K. Han, and Y. Wang. Generalized out-of-distribution detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1359–1381, 2021. doi: 10.1109/TPAMI.2021.3117323. URL https://arxiv.org/abs/2110.11334.

[110] P. Yaskov. A short introduction to random matrix theory. *arXiv preprint arXiv:1608.04850*, 2016.

[111] H. Yilmaz and S. Hooker. Ood detection for text via attention entropy and representation variance. *arXiv preprint arXiv:2208.10545*, 2022. URL https://arxiv.org/abs/2208.10545.

[112] Y. Yin, J. Li, C. Wang, and Y. Li. Mhaldet: Multimodal hallucination detection via attention entropy and cross-modal alignment. *arXiv preprint arXiv:2312.10457*, 2023. URL https://arxiv.org/abs/2312.10457.

[113] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1612.03928.

[114] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[115] P. Zhang, X. Li, X. Hu, et al. Vinvl: Making visual representations matter in vision-language models. In *Proceedings of CVPR*, 2021. URL https://arxiv.org/abs/2101.00529.

[116] W. Zhao. Enhancing user engagement and satisfaction through personalized news recommendation systems. *Applied and Computational Engineering*, 69:13–18, 2024. doi: 10.54254/2755-2721/69/20241454.

A | Appendix A

Portions of this thesis include material previously published as preprints on arXiv. The following works are included:

- **EigenTrack: Spectral Activation Feature Tracking for Hallucination and Out-of-Distribution Detection in LLMs and VLMs**, arXiv:2509.15735.
- **RMT-KD: Random Matrix Theoretic Causal Knowledge Distillation**, arXiv:2509.15724.

According to arXiv's reuse policy:

"If you are the copyright holder of the work, you do not need arXiv's permission to reuse the full text." (Source: `https://info.arxiv.org/help/license/reuse.html`)

I am the copyright holder of these works and therefore retain the right to reuse their content in this thesis.

List of Figures

List of Tables

List of Abbreviations

Acronym	Meaning
BBP	Baik–Ben Arous–Péché
CNN	Convolutional Neural Network
FNN	Feed Forward Neural Network
GRU	Gated Recurrent Unit
KD	Knowledge Distillation
KL	Kullback–Leibler
LLM	Large Language Model
LSTM	Long Short Term Memory
MP	Marčenko–Pastur
OOD	Out of Distribution
RMT	Random Matrix Theory
RNN	Recurrent Neural Network
TW	Tracy–Widom

Acknowledgements

I want to thank my family for always being there for me, supporting me from the very beginning of university, and never letting me lose sight of my goals. I also want to thank my companions in Mantova for all the moments of fun, laughter, and distraction that helped me take a break from stress and recharge. To my friends in Milan, thank you for being part of my everyday life, for the constant support throughout this journey, and for making those years unforgettable. To my group in Chicago and my roommates, thank you for sharing this adventure of living abroad during the double degree program at UIC Chicago, for facing the challenges and the excitement together, and for making this experience one I'll never forget. I would also like to thank professor Marco Brambilla for supporting my thesis defense at UIC Chicago, and for his support during the thesis development for my graduation at Politecnico di Milano, as well as for his guidance throughout this path. Finally, I want to thank my co-advisor, professor Amit Trivedi, for his mentorship, and all the people of AEON Lab for helping me give the best of myself.

DE