POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Computer Science Engineering - Ingegneria Informatica

Author: Davide Ettori

Advisor: Prof. Marco Brambilla

Co-advisor: Prof. Amit Ranjan Trivedi

Academic year: 2025-2026

1. Introduction

This thesis targets two linked challenges: reliability of large language and vision-language models (hallucinations and distribution shift) and efficiency at scale. Reliability failures erode trust, while the resource demands of large models limit deployment; a unified approach is therefore valuable. The thesis argues for principled, interpretable tools grounded in spectral geometry and Random Matrix Theory (RMT): eigenvalue spectra separate structure from noise, and the Marchenko–Pastur law plus the spiked covariance model formalize this separation into compact signatures of internal dynamics [1, 7, 8]. Related work on hallucination detection includes black-box output checks, gray-box logit/attention uncertainty, and white-box activation probes [6, 9]. For Out-of-Distribution (OOD) detection, entropy- and energy-based scoring of logits are common baselines, alongside representation-based and spectral methods [4, 10]. For efficiency, mainstream compression uses distillation, pruning, and quantization [2, 3, 5]. This thesis connect to these lines of work but emphasize spectral criteria that unify reliability and compression. A central motivation is that output-based or static uncertainty often misses failures that are visible in the evolving dynamics of internal representations. Spectral statistics provide compact, interpretable signals that can be tracked across layers and time, making them useful for both diagnosis and intervention. The first contribution, **EigenTrack**, detects hallucination and OOD behavior by tracking spectral descriptors over time with lightweight recurrent classifiers, enabling early warnings without modifying the base model. The second contribution, **RMT-KD**, applies the same spectral principles to compression, projecting onto outlier eigen-directions and self-distilling the reduced model to preserve accuracy while cutting size, latency, and energy.

2. Background

Random Matrix Theory (RMT) studies eigenvalue statistics of large random matrices (each entry is a random variable) and provides null models for high-dimensional noise. For symmetric i.i.d. random matrices, with variance σ^2, the Wigner semicircle law describes the limiting density, in which eigenvalues concentrate in $[-2\sigma, 2\sigma]$. For covariance-type matrices, the Marchenko–Pastur (MP) law gives the asymptotic spectrum. If $X \in \mathbb{R}^{n \times p}$ has i.i.d. ran-

dom entries with variance σ^2 and $p/n \to c$, then the eigenvalues of $C = \frac{1}{n}X^\top X$ lie in a bulk $[\lambda_-, \lambda_+]$ with $\lambda_\pm = \sigma^2(1 \pm \sqrt{c})^2$ and follow the distribution described by the MP law [7]. The MP bulk is a calibrated reference for noise-like activations. The Tracy–Widom distribution characterizes fluctuations of the largest eigenvalue at the upper spectral edge. A key framework is the spiked covariance model, where low-rank signal (spikes) is embedded in isotropic noise. The signal is: $\mathbf{\Sigma} = \sigma^2\mathbf{I} + \sum_{i=1}^{k} \theta_i u_i u_i^T$ where the first component is noise, θ_i are spike strengths and u_i are orthonormal signal directions. The Baik–Ben Arous–Péché (BBP) transition states that when a spike exceeds a threshold, $\theta > \sigma^2(1 + \sqrt{c})$, its sample eigenvalue detaches from the MP bulk and becomes an outlier [1]. In practice, those outliers indicate structured, task-relevant directions, while the bulk is noise-like with low energy eigenvalues.

3. Methodology: Eigentrack

EigenTrack is a real-time reliability monitor that attaches to a pretrained (Open Source) LLM or VLM and tracks how the geometry of hidden activations evolves during generation. The core hypothesis is that factual, in-distribution reasoning yields structured representations with a small number of dominant eigen-directions (consistent with the spiked covariance model), while hallucinations and OOD drift push the spectrum toward noise-like behavior (consistent with RMT), enabling the distinction. RMT provides the guiding principle: the MP bulk separates noise from informative structure, and deviations from this baseline become reliable failure indicators [1, 7]. Rather than relying on output probabilities, EigenTrack monitors the evolution of internal dynamics and therefore can flag risk early, before the model commits to a long hallucinated continuation. At each decoding step, EigenTrack collects hidden activations from a subset of layers and concatenate them into an activation vector. The last N activation vectors are stacked into a matrix representing a sliding window of recent states. Singular Value Decomposition (SVD) yields the eigenvalues of the windowed covariance, from which a compact descriptor vector is computed, Figure 1. The most informative descriptors include spectral entropy (dispersion), leading-eigenvalues mass (concentration), eigengaps (ratio between subsequent eigenvalues), and divergence (KL Divergence and Wasserstein distance) from the MP baseline. These descriptors are interpretable: they quantify whether the representation is collapsing toward isotropic noise or retaining structured, low-rank directions. Importantly, EigenTrack does not require access to gradients, training data, or changes to the base model; it is a monitoring head that can be attached post hoc and run concurrently with inference.

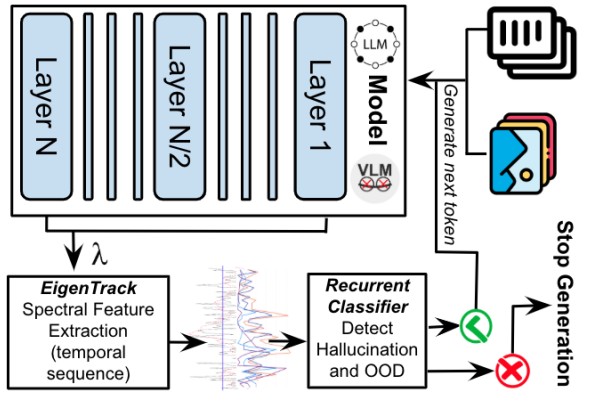

Figure 1: General architecture of EigenTrack: spectral features extracted from hidden activations are streamed into a recurrent discrepancy detector, which outputs early warnings.

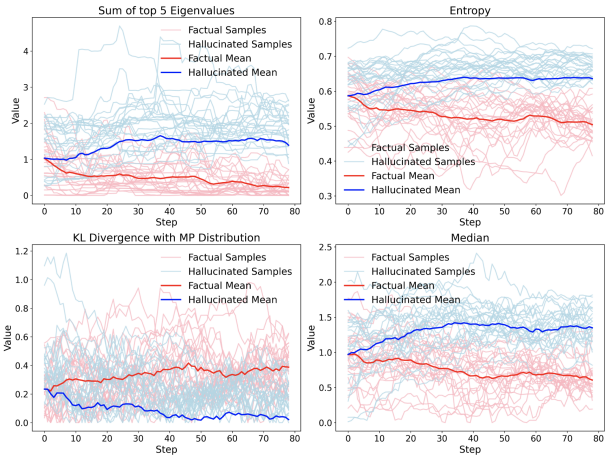

Figure 2: Temporal evolution of spectral statistics: hallucinated sequences tend to stay closer to the MP-like regime, while factual sequences diverge toward a more structured spectra.

Single spectral snapshots can be ambiguous, so EigenTrack models the time series of descriptors with a lightweight recurrent head (RNN/GRU/LSTM). This head learns characteristic trajectories of stable versus unstable generations while remaining computationally negligible compared to the base model. The out-

put is a per-step risk score that can trigger early intervention before hallucinated content is fully produced. The monitoring is non-invasive: the base model is unchanged and only a subset of layers is sampled, keeping overhead low and latency bounded. In practice, sampling one layer every few transformer blocks is sufficient because adjacent layers exhibit highly correlated spectral behavior. This design choice yields a favorable accuracy–latency trade-off and makes EigenTrack suitable for online deployment. From a systems perspective, EigenTrack adds minimal overhead: covariance computation uses short windows and truncated eigensolvers, the recurrent head adds only a few thousand parameters, and updates are constant per token.

Experimental Setup: EigenTrack is evaluated on open-source LLMs and VLMs (LLaMa, Qwen, Mistral, LLaVa) for hallucination and OOD detection. Hallucinations are induced via a controlled QA pipeline, where context-question pairs are sampled from HotPotQA and questions are randomly swapped with unanswerable ones (where the answer does not appear in the context) 50% of the time, and a larger LLM-as-a-judge model (e.g. LLaMa 8B) classifies whether the smaller model hallucinated. OOD evaluation contrasts in-distribution data from WebQuestions (common in web-sourced training corpora) against OOD samples from EurLex (domain-specific legal text from EU parliament).

Results: Across model families, EigenTrack consistently achieves strong AUROC for hallucination detection, with GRU heads typically outperforming RNNs and LSTMs. Performance improves with model scale, suggesting that larger models exhibit richer spectral signatures that are easier to separate. Figure 3a summarizes hallucination detection across architectures and recurrent heads. Additionally, EigenTrack achieves strong OOD performance across the same model families. Figure 3b shows consistent AUROC that mirror the hallucination trends, indicating that the spectral-temporal features generalize to broader distribution shifts. Beyond aggregate performance, the temporal plots in Figure 2 show that hallucinated generations follow distinct spectral trajectories: entropy remains higher, KL divergence from the MP baseline stays lower, and eigengaps narrow, consistent with a drift toward noise-like activations.

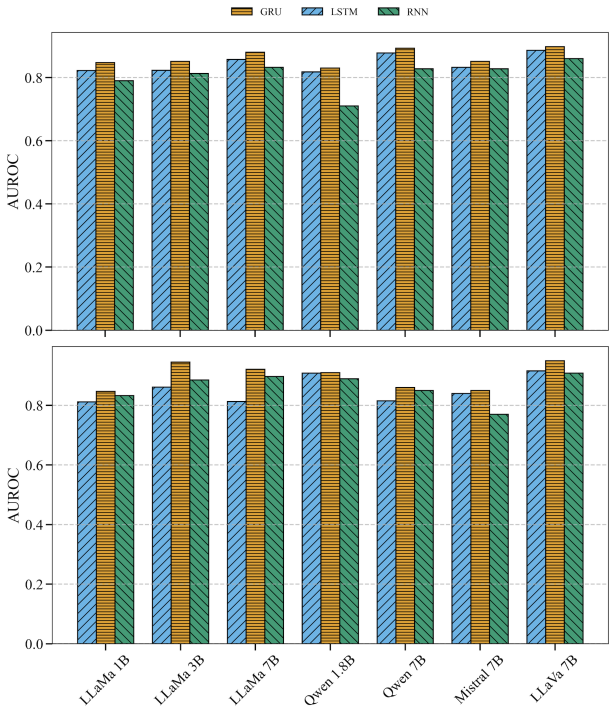

Figure 3: Hallucination **(a, top)** and OOD **(b, bottom)** detection metics across models and classifiers. Sliding window length is 30 tokens.

Table 1: AUROC comparison on LLaMa for hallucination detection on HotPotQA data.

Method	LLaMa1B	LLaMa3B	LLaMa7B
EigenTrack	**0.842**	**0.861**	**0.894**
LapEigvals	0.785	0.819	<u>0.871</u>
INSIDE	0.753	<u>0.831</u>	0.810
SelfCheckGPT	0.739	0.804	0.809
HaloScope	<u>0.820</u>	0.827	0.861

These dynamics explain why temporal modeling is crucial: risk emerges as a gradual spectral drift rather than a single anomalous point. Overall, EigenTrack indicates that spectral dynamics provide early, interpretable failure signals while remaining lightweight, supporting the thesis claim that RMT-based geometry can ground practical reliability tools. Table 1 summarizes state-of-the-art hallucination detection on the LLaMa family, with additional comparisons mentioned in the thesis document.

Ablation Study: Figure 4a shows AUROC and inference latency versus sliding-window length. Short windows yield higher AUROC by capturing fine-grained dynamics but increase latency due to more evaluations; longer windows reduce latency while gradually degrading performance.

A clear knee at 25 tokens offers the best trade-off. Below this range, spectral estimates are noisy and latency rise; above it, excess context oversmooths signals and degrades detection, indicating that hallucination dynamics lie within a moderate temporal horizon. Figure 4b plots AUROC against the number of observed tokens. Performance rises sharply from chance within the first few tokens and quickly saturates, showing that hallucination cues emerge early, with diminishing gains from later tokens, supporting early stopping to balance accuracy and latency. This indicates that hallucination isn't a sudden event and it can be detected, from internal activations, before manifesting in the text output.

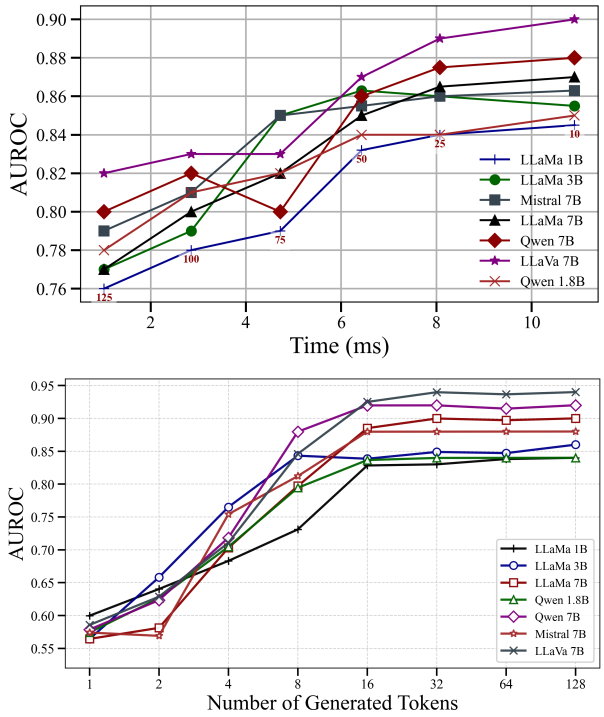

Figure 4: AUROC and latency as a function of sequence window length of GRU on hallucination dataset **(a, top)** and AUROC as a function of total generated tokens **(b, bottom)**.

4. Methodology: RMT-KD

RMT-KD is a compression framework that uses RMT to identify and preserve only the causal directions of hidden activations, yielding smaller models without sacrificing accuracy. The idea is that activation spectra exhibit a noise bulk predicted by the MP law, while task-relevant structure appears as outlier eigenvalues beyond the bulk edge; those outliers define the subspace worth keeping [1, 7]. RMT-KD, shown in Figure 5, operationalizes this into a staged procedure: train until a target level of accuracy is achieved (so we know there is signal in the model), analyze first layer activations on a calibration subset, estimate the upper MP bulk edge, project onto the outlier eigenvectors subspace, self-distill to recover accuracy and repeat for each layer until target compression is met.

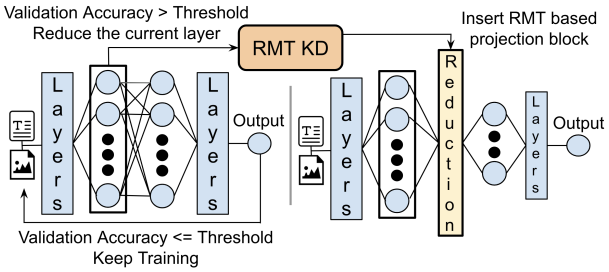

Figure 5: Overview of the iterative RMT-KD pipeline: spectral analysis estimates the MP bulk edge, outlier eigenvectors define the causal subspace, and self-distillation stabilizes training.

At each reduction step, activations from a target layer are collected to form a covariance matrix. The eigenvalue spectrum is matched to the MP density to estimate the noise variance and its upper edge λ_+. Eigenvalues above λ_+ are treated as signal, and their eigenvectors define a projection that reduces the layer width while preserving the most informative directions, as shown in Figure 6. Because the projection is dense, the resulting model remains hardware-friendly and does not require sparse kernels.

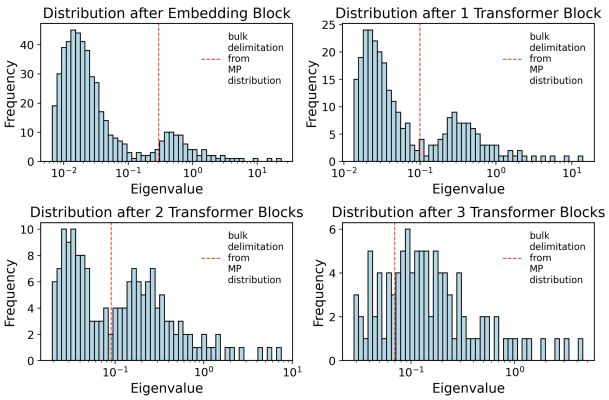

Figure 6: Evolution of the empirical eigenvalue distribution of activation covariances across layers of BERT on SST data during training.

Self-distillation is the stabilizing step: after each projection, the reduced model (student) is fine-tuned to match the logits of the pre-reduction checkpoint (teacher) while still optimizing the

task loss. This "teacher-from-previous-stage" scheme transfers decision boundaries into the smaller subspace, preventing catastrophic forgetting and allowing aggressive reductions to accumulate layer by layer. The process is modular and can be applied to transformers and CNNs with minimal architectural modifications.

Experimental Setup: RMT-KD is evaluated on BERT-base and BERT-tiny across GLUE tasks (SST, QQP, QNLI) and on ResNet-50 for CIFAR-10. Models share training settings; only the RMT projections and self-distillation schedule vary. A small calibration subset is used for spectral estimation, and performance is tracked in accuracy, parameter reduction, throughput, power, memory, and energy per inference.

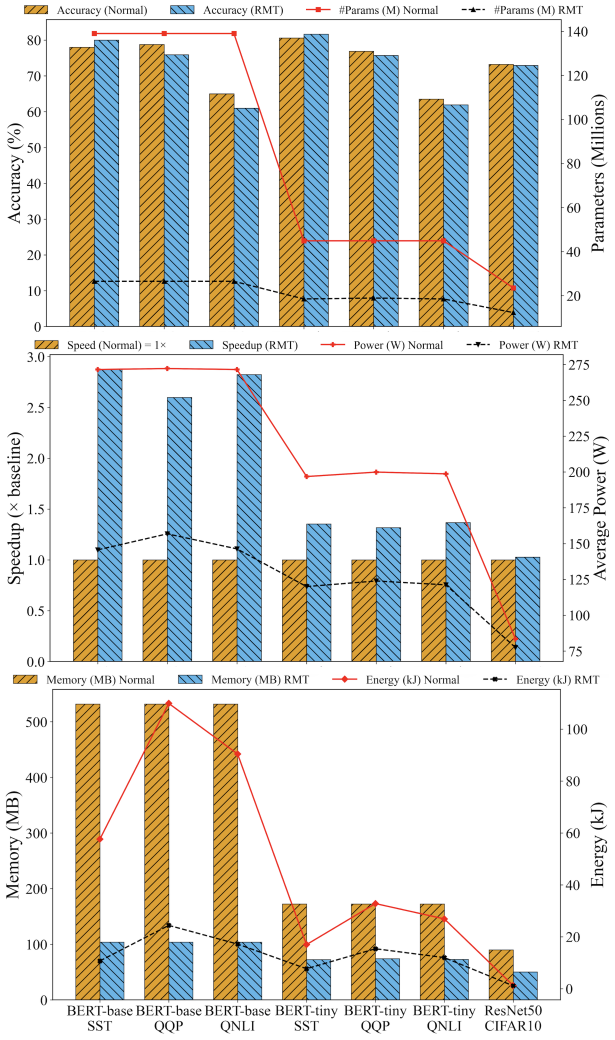

Figure 7: Performance metrics of RMT-KD models relative to baselines: **(a, top)** accuracy and parameter reduction, **(b, middle)** inference speedup and power reduction, **(c, bottom)** memory footprint and energy per inference.

Results: RMT-KD achieves large parameter reductions while maintaining or improving accuracy. Figure 7a shows the accuracy–compression trade-off: BERT-base sees reductions around 80% with slight accuracy gains on SST/QQP, BERT-tiny retains accuracy with 60% reduction, and ResNet-50 achieves nearly 50% reduction with minimal loss. The pattern suggests that spectral filtering removes redundant or noisy directions and can act as a regularizer, improving generalization even as capacity shrinks. The gains are consistent across architectures, indicating that the spectral criterion captures a general property of learned representations rather than a task-specific artifact. Efficiency gains are reflected in Figure 7b: throughput increases (nearly 3× on BERT-base in some tasks) while power consumption drops, indicating real hardware-level benefits from dimensionality reduction. Figure 7c complements this view by showing that memory footprint and total energy per inference decrease substantially, especially for larger models. These results matter for deployment because they reduce both latency and operational cost, and they do so without introducing sparsity; the compressed models remain dense and therefore align well with standard GPU kernels. This confirms that spectral projection is a principled alternative to heuristic pruning or low-rank factorization, preserving dense computation while cutting cost.

Table 2: Compression ratio (Red.) and accuracy change (Acc.) comparison with other methods.

	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 2 provides a compact state-of-the-art comparison against representative distillation baselines (data for BERT are the average performance between SST, QQP and QNLI, while CIFAR10 is used for ResNet); additional metrics and comparisons are reported in the thesis. The takeaway is that RMT-KD achieves both stronger compression and competitive performance, remaining dense and hardware-friendly.

Ablation Study: Compression aggressiveness is controlled by the quantile used to initialize the MP law variance. That is the expected variance in the activations matrix data, estimated as a quantile of the empirical eigenvalue distribution of the covariance matrix. Figure 8 shows that moderate quantiles (around the median) yield the best accuracy–compression trade-off, while higher quantiles become overly aggressive and degrade accuracy. This provides a tunable parameter balancing efficiency and performance. ResNet offers greater reduction potential, while BERT is constrained by fixed embedding and token projection layers, which dominate beyond 40% quantile and limit further compression.

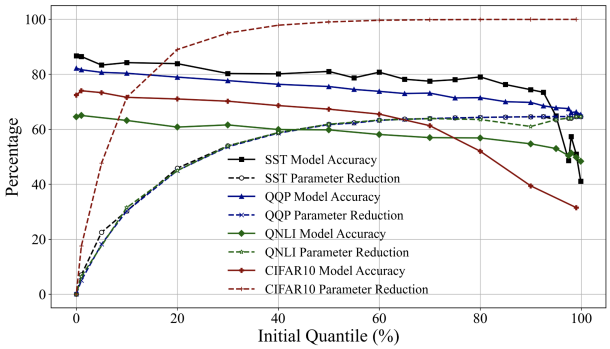

Figure 8: Effect of the variance initialization quantile on the compression-accuracy tradeoff.

5. Conclusions

This thesis shows that spectral geometry and RMT offer a unified view for improving both reliability and efficiency in deep learning. EigenTrack provides early, interpretable warnings of hallucination and OOD behavior without changing the base model. RMT-KD preserves causal eigen-directions via dense projection and self-distillation, achieving strong compression with minimal accuracy loss and clear system-level gains. These contributions position eigenvalue dynamics as a shared language for diagnosis and optimization. Limitations include evaluation scope and the cost of spectral computation for very large layers. Future work should scale to larger multimodal models, explore hybrid systems combining RMT analysis on attention matrices, and integrate approximate eigensolvers.

References

[1] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[2] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021.

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

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

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[6] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*, 2023.

[7] Vladimir A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

[8] Charles H. Martin and Michael 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.

[9] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *ICML*, 2023.

[10] Haoran Sun, Zhaoning Wang, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.