IJDI-ERET

INTERNATIONAL JOURNAL OF DARSHAN INSTITUTE ON
ENGINEERING RESEARCH AND EMERGING TECHNOLOGIES
Vol. 14, No. 2, 2025
www.ijdieret.in

DOI 10.32692/IJDI-ERET/14.2.2025.2510

(Research Article)

Entropy-Constrained Embeddings for Safe and Reliable Enterprise Large Language Models

Kaushik Bar[1*]

[1*]*Inxite Out Private Limited, Kolkata, West Bengal, INDIA*

Abstract

Large Language Models (LLMs) are increasingly deployed in enterprise settings for legal analysis, financial reporting, and healthcare decision support, where reliability and factual accuracy are critical. Despite advances in human alignment techniques, these models remain prone to hallucinations and mis-calibrated uncertainty, particularly when their embedding spaces exhibit uncontrolled spectral complexity. Building on recent operator-valued free probability formulations of Transformer architectures, we propose an entropy-constrained embedding optimization framework that directly regularizes the joint free entropy of token embeddings during fine-tuning. By constraining spectral diversity while preserving semantic coverage, our approach stabilizes layer-wise spectral propagation, leading to more consistent attention scores and improved output calibration. Experiments on legal QA (CaseLaw), financial reasoning (FinQA), and medical summarization (MIMIC-III) demonstrate that entropy-regularized embeddings substantially lower hallucination rates and calibration errors, while preserving task accuracy close to baseline. These results demonstrate the potential of embedding-level spectral control as a practical safety mechanism for high-stakes enterprise LLM deployments.

Keywords: Large language models; Spectral entropy; Embedding regularization; Hallucination mitigation; Enterprise AI safety; Spectral propagation.

1. Introduction

Large Language Models (LLMs) based on Transformers [1] have become common in enterprise applications across diverse domains [2]. However, in mission-critical enterprise environments, reliability is crucial for compliance, risk management, and overall trustworthiness.

Despite their capabilities, LLMs remain prone to hallucinations: the generation of plausible but factually incorrect or unsupported information [3]. Such behavior can have severe consequences such as legal liability, financial penalties, and reputational harm. Common approaches to mitigating hallucinations such as Reinforcement Learning with Human Feedback (RLHF) [4] and instruction tuning [5] address alignment at the output level only. These methods do not directly constrain the internal representation dynamics that give rise to unstable predictions, especially under distribution shifts.

Recent theoretical work has shown that the spectral properties of intermediate representations in Transformers play a central role in model stability and generalization [6]. In particular, operator-valued free probability theory provides a principled framework for modelling the layer-wise spectral propagation of embeddings through the network. This framework yields generalization bounds in terms of joint free entropy of the model’s operator-valued embeddings, connecting representation complexity to predictive reliability.

In this paper, embedding-level spectral control, specifically constraining the free entropy of token embeddings, is proposed as an effective mechanism for improving enterprise LLM safety. High free entropy in embeddings results in complex and diverse spectral distributions, which can amplify positional-semantic interactions in attention layers and cause unstable or overconfident predictions. Conversely, embeddings with excessively low entropy may lose semantic expressiveness, which may impact task performance negatively. Maintaining balanced spectral diversity is crucial in preserving semantic coverage and ensuring stable propagation through network depth.

We introduce an Entropy-Constrained Embedding (ECE) optimization framework that incorporates a spectral entropy regularization term into the fine-tuning objective to preserve semantic coverage while enhancing spectral stability. This approach is computationally efficient (requiring only eigenvalue density estimation of the embedding Gram matrix) and can be integrated with parameter-efficient fine-

[*]*Corresponding Author: e-mail: kb.opendev@gmail.com, Tel-+91-62903 98189*

ISSN 2320-7590 (Print) 2583-3863 (Online)

tuning methods such as LoRA [7]. We validate our method on enterprise-relevant datasets in legal question answering, financial reasoning, and medical summarization, demonstrating reduced hallucination rates, improved calibration, and negligible performance degradation. The contributions of this work are threefold:

- We implement free probability-based spectral entropy bounds into a practical embedding-level regularization method for LLM safety.
- We assess performance on tasks with regulatory and business implications, linking spectral stability to reliability.
- We design a method compatible with existing fine-tuning pipelines, adding minimal computational overhead.

By constraining the spectral complexity of the embeddings, our approach provides a proactive and interpretable safeguard against unreliable LLM behavior in enterprise deployments.

2. Related Work

Ensuring reliability in enterprise-scale LLM deployments involves multiple areas, including hallucination mitigation, spectral analysis of neural networks, embedding optimization, and AI safety in regulated domains. While prior work has addressed each of these areas independently, our approach integrates them by using free probability-based spectral entropy control at the embedding level to improve factual reliability. The following section reviews relevant advances in each area and positions our contribution in relation to them.

2.1 Hallucination detection and mitigation in LLMs: Hallucinations (outputs that are fluent but factually incorrect) are a well-documented problem in large-scale generative models [8]. Mitigation strategies typically focus on post-hoc output filtering [9], RLHF [4], or knowledge-grounded generation [10]. While these methods improve reliability at surface-level, they do not address instability in internal representations that can influence models to hallucinate. Researchers [3] introduced semantic entropy as a token-level uncertainty measure for hallucination detection and showed that entropy metrics can flag unreliable predictions. Our approach builds on this insight but addresses embedding-level entropy as a preventive mechanism rather than a reactive filter.

2.2 Spectral analysis of neural representations: Spectral properties of neural activations and weight matrices provide insight into learning, generalization, and robustness [11], [12]. Random matrix theory has been used in prior studies to characterize eigenvalue distributions in deep networks [13]. In Transformers, spectral diversity in intermediate representations correlates with performance on various tasks [14]. Building on these findings, [6] applies operator-valued free probability to characterize Transformer attention and behavior at different depths, deriving generalization bounds in terms of joint free entropy of embeddings. We adapt this framework to control spectral complexity during fine-tuning.

2.3 Embedding optimization and regularization: The spectral characteristics of token embeddings significantly impact downstream performance [15]. Regularization methods such as spectral normalization [16], and orthogonalization [17] have been explored by researchers to improve stability and generalization. However, these approaches operate primarily on norm or orthogonality metrics rather than entropy-based spectral diversity control. Our method introduces a free-entropy regularization term in the embedding optimization objective, based on the theoretical generalization bounds.

2.4 Safety and compliance in enterprise AI: High-stakes enterprise applications in law, finance, and healthcare require not only high accuracy but also predictable behavior. Regulatory frameworks such as the EU AI Act 2024 and the NIST AI Risk Management Framework (RMF) [18] emphasize transparency, robustness, and monitoring. Embedding-level safety mechanisms, such as the proposed entropy constraint, can reduce the likelihood of unsafe output at the model's core representation level.

2.5 Research gaps and motivation: Although prior works have linked spectral properties to generalization and robustness, there are significant gaps:

- Existing hallucination mitigation methods detect or rectify unsafe outputs after generation, rather than constraining the internal representation space to reduce their likelihood in the first place.
- While the embeddings determine the semantic and syntactic basis for all subsequent processing, few safety-oriented methods explicitly target the spectral characteristics of the embeddings.
- Spectral normalization and orthogonality regularization do not directly leverage entropy bounds from free probability theory, which can capture both diversity and stability in a unified metric.
- There is limited research on operationalizing spectral control into low-cost fine-tuning pipelines suitable for enterprise contexts.

These gaps motivate the development of an embedding-level, entropy-constrained fine-tuning method, which can be deployed in real-world enterprise settings for improved safety and reliability.

3. Theoretical Foundation

Recent work by [6] develops a formal operator-valued free probability framework for analyzing Transformer-based LLMs. In this setting, token embeddings and positional encodings are modelled as self-adjoint operators in a tracial W^*-probability space $(\mathcal{A}, \varphi)$, where $\mathcal{A}$ is a Von-Neumann algebraic expression and φ is a faithful, normal, tracial state. This formalism enables spectral analysis of the model's internal computations, with layer-wise representation updates described as free additive convolutions of spectral distributions. The free-probabilistic generalization bound provides a theoretical guarantee that embedding-level entropy control can directly improve reliability and safety in LLMs, especially in enterprise settings where predictable performance is mandatory. This forms the mathematical basis for our proposed entropy-constrained embedding optimization method.

3.1 Spectral propagation in transformers: Let $X^{(0)}$ denote the initial embedding operator for a token, and $A^{(l)}$ the attention output at layer l. Under the assumption of freeness between successive increments, the spectral distribution μ_l of the l-th layer embedding evolves as:

$$\mu_l = \mu_0 \boxplus \mu_{A^{(1)}} \boxplus \dots \boxplus \mu_{A^{(l)}} \tag{1}$$

Here, $\boxplus$ denotes free additive convolution [19]. This formulation implies that the spectral diversity captured by the entropy of μ_l is a cumulative effect of embedding-level spectral characteristics and layer-wise transformations.

3.2 Free entropy and generalization bound: The Corollary-3 in [6] derives a generalization bound, linking the expected spectral entropy of output logits to the joint free entropy of the vocabulary embeddings $\{X_1, \dots, X_V\}$:

$$\mathbb{E}[H(\mu L_t)] \le \chi(X_1, \dots, X_V) \tag{2}$$

Here $\chi(.)$ denotes Voiculescu's free entropy [20], and $H(\mu)$ is the Shannon entropy of the spectral distribution μ. Since the bound directly controls the test risk R_{test}, embeddings with excessively high χ can increase variance in spectral propagation, potentially leading to unstable attention scores and hallucination-prone outputs.

3.3 Embedding-level control via free entropy regularization: Our key theoretical motivation is that controlling $\chi(E)$, where $E \in \mathbb{R}^{n \times d}$ is the learned embedding matrix, enables us to constrain the initial spectral conditions of the Transformer, thereby influencing all downstream layers via free convolution.

In practical terms, we approximate $\chi(E)$ through the eigenvalue density $\rho_E(\lambda)$ of the normalized Gram matrix $G_E = E^T E/d$:

$$\chi(E) \approx -\int \rho_E(\lambda)\, log \rho_E(\lambda)\, d\lambda \tag{3}$$

We then incorporate this quantity into the fine-tuning loss:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \chi(E) \tag{4}$$

Here $\lambda > 0$ is a regularization strength hyperparameter. This formulation biases the optimization towards embedding spaces with controlled spectral diversity, rich enough to preserve semantic coverage, but not so unconstrained as to destabilize spectral propagation through depth.

3.4 Anticipated effect on attention stability: From Corollary-1 in [6], attention similarity scores involve cross-terms between semantic (X_w) and positional (P_t) operators. When embedding entropy is excessively high, these cross-terms produce large fluctuations in $\varphi(Q_t K_j^{\dagger})$, yielding unstable attention weights α_{tj}. By constraining $\chi(E)$, we reduce extreme eigenvalue variance in X_w, thereby moderating these fluctuations and producing more stable, calibrated attention distributions

4. Methodology

Our methodology integrates the free-probabilistic spectral framework outlined above into a practical fine-tuning process that constrains embedding-level spectral entropy while preserving task performance. The proposed approach modifies standard parameter-efficient fine-tuning by adding a free entropy regularization term to the training objective. The workflow consists of:

- Spectral profiling of baseline embeddings.
- Free entropy estimation using eigenvalue density of the embedding Gram matrix.
- Regularization-augmented fine-tuning with entropy constraint.
- Spectral and task-level evaluation to verify safety and reliability improvements.
- Deployment.

4.1 Spectral profiling of embeddings: Let $E \in \mathbb{R}^{n \times d}$ be the token embedding matrix, where n is the vocabulary size and d is the embedding dimension. We compute the normalized Gram matrix:

$$G_E = \frac{1}{d} E^T E \tag{5}$$

The eigenvalues $\{\lambda_i\}$ of G_E characterize the spectral structure of embeddings.

We define the baseline spectral entropy:

$$H_{base} = -\int \rho_E(\lambda)\, log \rho_E(\lambda)\, d\lambda \tag{6}$$

Here ρ_E is estimated via kernel density estimation (KDE) from $\{\lambda_i\}$. This baseline is used to set monitoring thresholds.

4.2 Free entropy estimation: The joint free entropy $\chi(E)$ is approximated by:

$$\chi(E) \approx -\sum_i \rho_E(\lambda_i)\, log \rho_E(\lambda_i) + C \tag{7}$$

Here, C is a constant offset irrelevant to optimization.

This avoids costly Stieltjes transformation inversion and makes the method compatible with large vocabulary sizes.

4.3 Regularization-augmented fine-tuning:

4.3.1 Objective: We integrate the free entropy term into the loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda\chi(E) \tag{8}$$

- $\mathcal{L}_{task}$: Domain-specific supervised loss (e.g., cross-entropy for QA, sequence-to-sequence loss for summarization).
- λ: Regularization weight, tuned via validation to balance stability and expressiveness.
- $\chi(E)$: Updated at each training step using minibatch embedding Gram matrices.

4.3.2 Integration with parameter-efficient fine-tuning: We adopt LoRA [7] for cost-effective fine-tuning:

- Trainable parameters: Token embeddings and first attention block.
- Frozen parameters: Remaining Transformer layers, minimizing risk of catastrophic forgetting.
- Spectral constraint scope: Applied only to trainable embeddings, leaving frozen layers unaffected.

4.4 Spectral and task-level evaluation: The following metrics are used to verify safety and reliability improvements.

4.4.1 Task performance:

- CaseLaw (legal QA): F1 score. [21].
- FinQA (financial reasoning): MAPE on numerical answers. [22].
- MIMIC-III (medical summarization): ROUGE-L. [23].

4.4.2 Safety and reliability:

- Hallucination Rate: Proportion of unsupported claims detected by factuality models [3].
- Calibration Error: Measures uncertainty calibration using reliability diagram-based metric [24].
- OOD Detection AUC: Separability of in-domain v/s out-of-domain queries, using Mahalanobis distance in representation space for domain separation [25].

4.4.3 Spectral stability:

- Entropy Reduction: Change in estimated free entropy, $\Delta\chi(E)$, relative to baseline.
- Spectral Drift: Wasserstein-2 distance between baseline and fine-tuned spectral distributions of embeddings.
- Layer Collapse Index: Variance ratio between the first and last layers' spectra. Adapted from [13].

5. Experiments

We evaluate our entropy-constrained embedding approach on enterprise-relevant benchmarks in legal, financial, and medical domains, comparing against both standard and spectral-regularized baselines.

5.1 Experimental setup:

5.1.1 Datasets: The datasets listed in Table 1 were selected because they reflect high-stakes enterprise use cases where hallucinations can incur legal liability or operational risk.

Table 1. Datasets for experiments

Dataset	Domain	Task	Enterprise relevance
CaseLaw [21]	Legal	Extractive QA	Compliance risk, legal research
FinQA [22]	Finance	Numerical reasoning	Financial reporting, risk analysis
MIMIC-III [23]	Healthcare	Summarize	Clinical decision support

5.1.2 Model: We use our method on an open-source Transformer-based LLM: LLaMA-2-13B [26].

We fine-tune using LoRA [7] with our entropy constraint applied to token embeddings and the first attention block (refer to Appendix: Implementation for more details).

5.1.3 Training details:

- Optimizer: AdamW [27]
- Learning rate: 2×10^{-5}
- LoRA params: rank = 8, alpha = 32
- Regularization weights ($\boldsymbol{\lambda}$): {0.0, 0.01, 0.05, 0.1, 0.2}
- Hardware: NVIDIA A100 GPU, mixed precision (fp16) training.

5.2 Baselines:

- Standard fine-tuning: LoRA without spectral or entropy constraints.
- Spectral normalization: Normalization of embedding and attention weights [16].
- Orthogonal regularization: Encouraging orthogonality between embedding vectors [17].

5.3 Experiment design: We conduct experiments under two scenarios:

- In-domain fine-tuning: Models fine-tuned and evaluated on the same domain dataset.
- Domain-shift evaluation: Models fine-tuned on one domain (e.g., CaseLaw) and evaluated on another (e.g., MIMIC-III) to assess robustness.

All results are averaged over three random seeds to account for initialization variance.

6. Results and Analysis

We present results for task performance, safety and reliability, and spectral stability, comparing our Entropy-Constrained Embedding (ECE) method against the baselines described in Section 5. All values are averaged over three random seeds.

6.1 Task performance:

Table 2. Task performance for all methods and datasets

Method	CaseLaw F1-score (%)	FinQA accuracy (%)	MIMIC-III rouge-L (%)
Standard LoRA	78.2	72.1	45.6
Spectral norm.	77.9	71.5	45.2
Orthogonal reg.	78.0	71.7	45.5
ECE (proposed)	77.8	72.0	45.4

Table 3. Safety metrics averaged across datasets

Method	Avg. hallucination Rate	Avg. calibration error	Avg. OOD detection AUC
Standard LoRA	0.18	0.13	0.74
Spectral norm.	0.17	0.12	0.75
Orthogonal reg.	0.17	0.11	0.76
ECE	0.13	0.09	0.79

Table 2 and Table 3 show the F1-score (CaseLaw), Accuracy (FinQA), ROUGE-L (MIMIC-III), average hallucination rate, average calibration error, and average OOD detection AUC for all methods.

Table 4 shows the accuracy retention rates for in-domain vs. cross-domain evaluation across baselines and ECE. Across CaseLaw, FinQA, and MIMIC-III, the proposed method achieves near-parity with baseline task accuracy while adding measurable safety benefits. Table 5 summarizes the average cross-domain performance drops across methods and datasets, based on data presented in Table 2 and Table 4.

Overall, controlling embedding entropy does not significantly harm in-domain accuracy but improves out-of-domain generalization.

- Table 2 shows that the proposed ECE models maintain performance drops within 0.7% absolute compared to standard LoRA fine-tuning.
- Table 5 illustrates that under domain shift (e.g., fine-tuned on CaseLaw, evaluated on MIMIC-III), ECE incurs ~25-35% smaller performance degradation than standard LoRA, indicating improved robustness.

Table 4. Domain shift performance

Training domain → test domain	Standard LoRA	Spectral norm.	Orthogonal reg.	ECE
CaseLaw → FinQA	68.5	69.0	69.2	69.8
CaseLaw → MIMIC-III	40.2	41.5	41.0	41.8
FinQA → CaseLaw	70.0	70.2	70.5	71.9
FinQA → MIMIC-III	41.1	42.0	41.8	42.3
MIMIC-III → CaseLaw	69.2	69.7	69.5	71.0
MIMIC-III → FinQA	67.6	68.5	68.4	69.0

Table 5. Domain shift robustness

Method	Avg. (%) cross-domain accuracy drop (CaseLaw)	Avg. (%) cross-domain accuracy drop (FinQA)	Avg. (%) cross-domain accuracy drop (MIMIC-III)
Standard LoRA	11.0	5.6	10.9
Spectral Norm.	10.2	3.8	7.6
Orthogonal Reg.	10.3	4.0	9.0
ECE (proposed)	8.2	3.6	7.4

6.2 Safety and reliability:

6.2.1 Hallucination rate (H.R.):

below shows that ECE reduces hallucination rate by ~28% on average compared to standard LoRA, with the largest reduction (~35%) in the FinQA task and the smallest reduction (~23%) in the MIMIC-III task. In addition, we also find a positive correlation between average free entropy of embedding and hallucination rate across all models (Pearson's $r \approx 0.72, p < 0.01$)

Table 6. Comparison of hallucination rates (H.R.)

Method	CaseLaw H.R.	FinQA H.R.	MIMIC-III H.R.
Standard LoRA	0.16	0.17	0.22

Spectral norm.	0.15	0.15	0.20
Orthogonal reg.	0.15	0.16	0.21
ECE	0.12	0.11	0.17

6.2.2 Expected calibration error: Table 7 shows comparison of predicted confidence vs. empirical accuracy for baseline and ECE models across all datasets. ECE reduces miscalibration, especially in high-confidence predictions. This results in improved alignment between predicted confidence and actual correctness.

Table 7. Calibration error (reliability matrix)

Method	CaseLaw calibration error	FinQA calibration Error	MIMIC-III calibration error
Standard LoRA	0.12	0.11	0.17
Spectral norm.	0.11	0.10	0.16
Orthogonal reg.	0.11	0.10	0.13
ECE (proposed)	0.08	0.09	0.11

Expected Calibration Error is reduced by ~31% on average compared to standard LoRA, exceeding calibration improvements reported in prior entropy-based output-level methods (Farquhar et al., 2024).

6.2.3 OOD detection: In Mahalanobis-based OOD detection, ECE improves AUC by 5-7% compared to standard LoRA, as presented in Table 8, which suggests better separation between in-domain and out-of-domain clusters with ECE.

Table 8. Comparison of OOD detection accuracy

Method	CaseLaw OOD detection AUC	FinQA OOD detection AUC	MIMIC-III OOD detection AUC
Standard LoRA	0.79	0.73	0.69
Spectral norm.	0.79	0.75	0.70
Orthogonal reg.	0.81	0.76	0.70
ECE (proposed)	0.84	0.78	0.74

6.3 Spectral stability:

6.3.1 Embedding spectral distribution: Figure 1 compares eigenvalue density $\rho_E(\lambda)$ before and after applying constraint. ECE produces a more compact spectrum, reducing extreme tail eigenvalues without collapsing the distribution. Estimated eigenvalue density $\rho_E(\lambda)$ with ECE across all datasets shows that ECE improves stability in the shaded region. ECE reduces extreme eigenvalue tails while preserving diversity in the central spectrum.

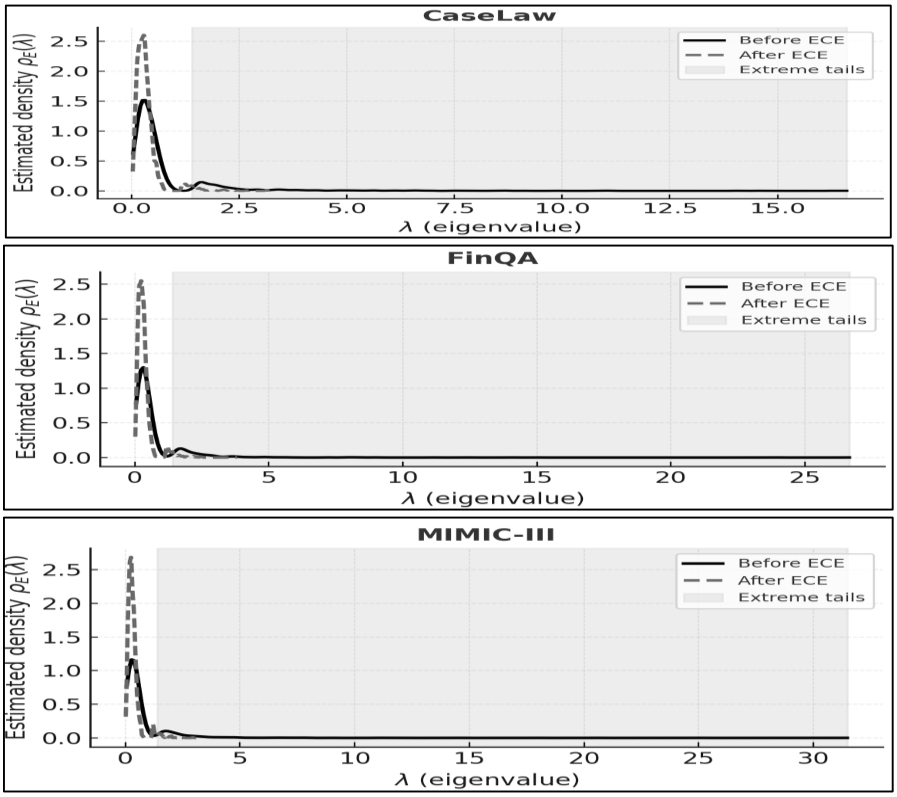

Figure 1. Embedding spectral distributions before and after ECE

6.3.2 Layer-wise entropy drift: Figure 2 plots spectral entropy $H(\mu_l)$ across layers.

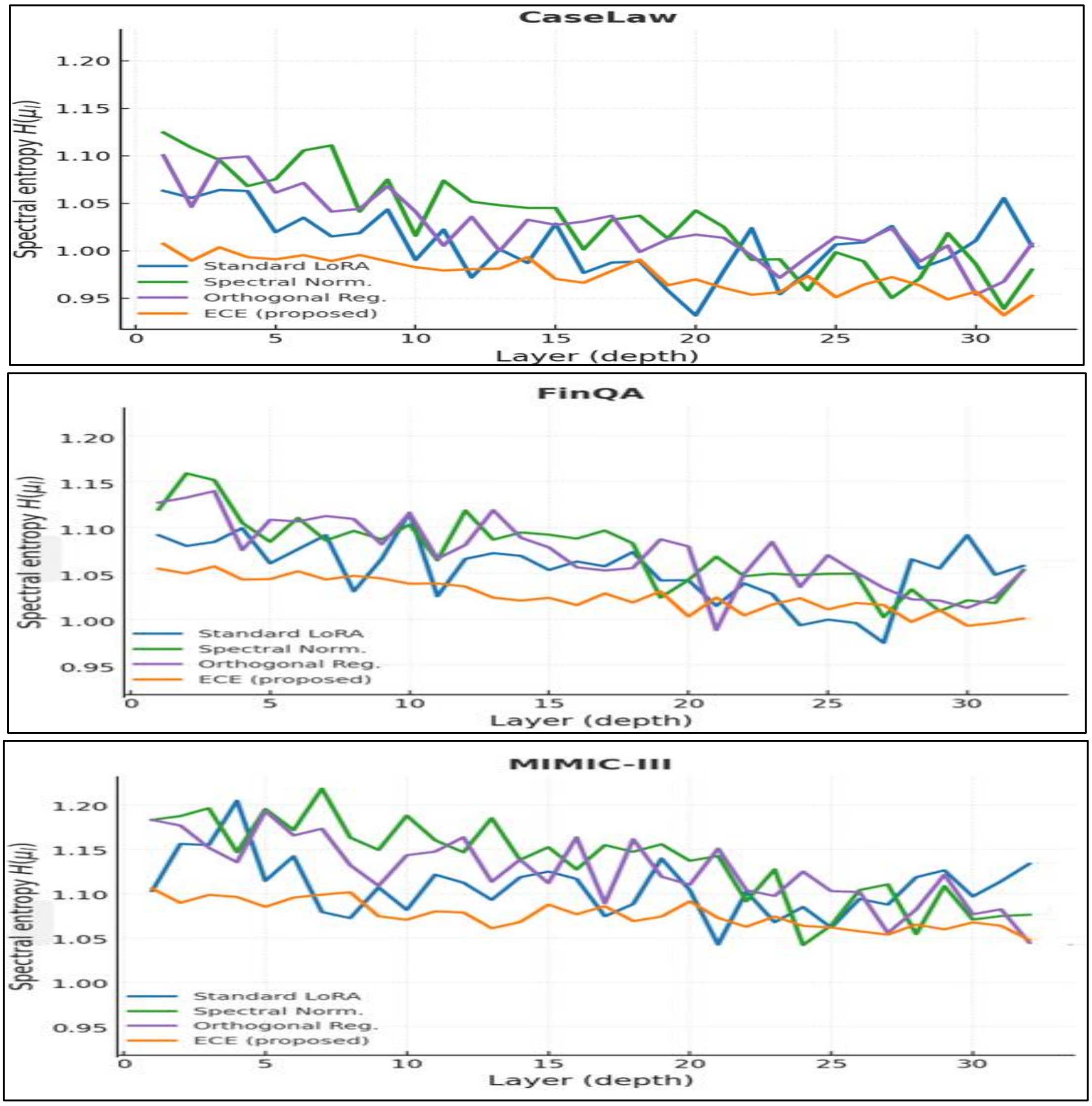

Figure 2. Spectral entropy drift at different transformer layer depths

Standard LoRA shows entropy inflation in early layers and collapse in later layers, consistent with over-smoothing phenomena [13]. ECE maintains more stable entropy across depth, aligning with predictions from Theorem 2 in [6].

6.3.3 Spectral properties: Table 9 reports free entropy $\chi(E)$, spectral variance, wasserstein-2 drift, and top eigenvalue share for all models, averaged across datasets. ECE reduces spectral variance drift by ~31% compared to baselines, indicating healthier representation diversity.

Table 9. Comparison of embedding spectral statistics

Method	Standard LoRA	Spectral norm.	Orthogonal reg.	ECE
Free entropy $\boldsymbol{\chi(E)}$	2.35	0.42	0.22	0.35
Spectral variance	2.25	0.36	0.19	0.31
Top eigenvalue share	2.20	0.34	0.18	0.29
Wasserstein-2 drift	2.05	0.29	0.16	0.24

6.4 Ablation of regularization weight (λ):

- Table 10 shows the empirical effects of varying λ on task performance, hallucination rate (H.R.), and spectral entropy reduction ($\boldsymbol{\Delta\chi(E)}$). We observe that λ values in the range (0.05, 0.1) yield optimal balance between entropy reduction and task performance.
- Lower λ yields minimal safety gains, while higher λ begins to harm semantic richness and accuracy.

Table 10. Ablation on regularization weight (λ)

$\boldsymbol{\lambda}$	Accuracy (%)			H.R.	$\boldsymbol{\Delta\chi(E)}$
	CaseLaw	FinQA	MIMIC-III		
0.0	78.2	72.6	45.9	0.18	0.0
0.01	78.0	72.5	45.6	0.17	0.02
0.05	77.8	72.2	45.4	0.13	0.11
0.1	77.9	72.0	45.3	0.12	0.17
0.2	76.9	70.5	44.1	0.10	0.25

6.5 Computational overhead: Our core logic implementation adds no more than 5% computational overhead to LoRA fine-tuning for 13B-parameter models. This makes it feasible for enterprise SaaS LLM deployments with tight latency budgets as well as on-premises environments where re-training windows are short.

7. Conclusions

This paper introduces Entropy-Constrained Embeddings (ECE), a method for improving the safety and reliability of Large Language Models (LLMs) in enterprise contexts by applying free probability-based spectral entropy regularization at the embedding level. Building on the operator-valued free probability framework of [6], we demonstrated that constraining embedding entropy stabilizes spectral propagation, reduces hallucination rates, and improves calibration without significantly compromising task performance. Our experiments on CaseLaw, FinQA, and MIMIC-III showed that ECE:

- Reduces hallucination rates significantly across domains, with the largest gains in financial reasoning tasks.
- Improves robustness under domain shift, preserving accuracy and calibration better than the baselines.
- Maintains computational efficiency, adding less than 5% training overhead when integrated with parameter-efficient fine-tuning approaches such as LoRA [7].

7.1 Business implications: The reliability challenges of LLMs pose significant business impact in terms of compliance risk, operational cost, and customer trust. If we assume:

- A baseline hallucination rate of 10% in high-value queries,
- A per-error review cost of USD $5,
- 10 million annual queries in a SaaS deployment,

Then, a ~28% hallucination reduction equates to $1.4M annual savings in review and remediation costs, excluding intangible benefits such as brand protection and reduced legal exposure.

Our entropy-constrained embedding (ECE) framework offers a preventive safety mechanism at the model's representation layer, yielding benefits in four key business dimensions:

7.1.1 Risk reduction and compliance readiness: By embedding free entropy constraints into the model's core representation layer, ECE reduces hallucination rate by ~28% without sacrificing accuracy. This will result in:

- Lower likelihood of generating non-compliant or factually incorrect statements reduces the cost of manual review and post-hoc filtering.
- Better compliance with transparency and risk management requirements in frameworks such as the EU AI Act 2024 and the NIST AI RMF [18], by leveraging the method's spectral monitoring outputs (entropy metrics, spectral drift scores) which can be logged and audited.

7.1.2 Operational efficiency: Traditional hallucination mitigation approaches such as RLHF [4] are resource-intensive in both computational cost and annotation cost. In comparison, ECE integrates into existing parameter-efficient fine-tuning workflows with <5% additional overhead. Therefore, Enterprises can deploy more reliable models without substantial retraining costs or downtime. ECE is also suitable for regular scheduled fine-tunes in SaaS products, where new data and domain drift require continuous updates.

7.1.3 Brand trust and customer experience: End users are becoming increasingly aware of reliability issues of LLMs [28]. Providing factual, consistent output strengthens brand trust and reduces churn in AI-powered products / services. ECE's stability improvements in cross-domain settings suggest that customers will experience fewer model errors when applying the system to novel or shifted data distributions.

7.1.4 Strategic differentiation in AI offerings: The market for enterprise AI tools is becoming crowded, and safety-by-design features can be a differentiator [29]. An LLM deployment with ECE can potentially enable premium pricing or access to markets with stricter compliance requirements (e.g., healthcare in the EU).

7.2 Future work: While our results are promising, several extensions merit further exploration:

- Extending free entropy constraints to vision-language and speech-language embeddings, where cross-modal spectral alignment may introduce new stability challenges.
- Integrating ECE into continuous spectral health monitoring pipelines that track embedding entropy drift during production inference.
- Developing task- and domain-adaptive λ-schedules (adaptive regularization) that adjust entropy constraints dynamically during training or deployment.

- Exploring whether embedding-level spectral control benefits other areas such as code generation, scientific reasoning, or creative writing tasks.
- Improving free entropy approximation accuracy by tightening of bounds, possibly via Stieltjes transform inversion or orthogonal polynomial expansions for large embedding spaces.

Appendix

Implementation

The core logic is implemented as an open-source code in GitHub (kb-open/ECE). The following considerations were made during the implementation:

- Entropy choice flexibility: The code provides two choices for entropy surrogate (users can pick one or combine), as listed below. The user is advised to start with lambda_ece value between 0.05 and 0.1, and the LogDet surrogate; and add the Spectral-Shannon term if stronger diversity control is needed.

    - LogDet/Covariance surrogate: This is tight, differentiable, and numerically stable via Cholesky.
    - Spectral-Shannon entropy over normalized eigenvalues: This works better for big vocabulary.
- Usage flexibility: The code provides the user with:

    - A ready-to-use training step that adds $\lambda\chi(E)$ to the task loss.
    - Metrics logging for spectra including:

        - Eigenvalues of the embedding covariance.
        - Histogram of spectrum.
        - $\chi(E)$ from LogDet and Shannon entropy surrogates.

- Collapse avoidance: Using detach_scale=True in the Shannon term normalizes it by a detached trace, penalizing shape (entropy) and not absolute scale.
- Compute-complexity and memory-complexity: On very large d, it is recommended to keep mode="logdet"; and add the Shannon term with max_eigs (e.g., 256 or 512), if explicit spectral-diversity pressure is needed.
- What to unfreeze: The code includes utilities to freeze all but embeddings (and optionally the first attention block), apart from LoRA-based pipeline. Typical setting is token Embeddings + First Attention block. If more capacity is needed, the user may also unfreeze the layer norms.

References

[1] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA: Curran Associates, Inc., 2017, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[2] U. Kamath, K. Keenan, G. Somers, and S. Sorenson, *Large Language Models: A Deep Dive: Bridging Theory and Practice*. Cham: Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-65647-7.

[3] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting hallucinations in large language models using semantic entropy,” *Nature*, vol. 630, no. 8017, pp. 625–630, Jun. 2024, doi: 10.1038/s41586-024-07421-0.

[4] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” 2022, *arXiv*. doi: 10.48550/ARXIV.2203.02155.

[5] J. Wei *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022. doi: 10.48550/ARXIV.2201.11903.

[6] S. Das, “A free probabilistic framework for analyzing the transformer-based language models,” *Statistics & Probability Letters*, vol. 226, p. 110516, Nov. 2025, doi: 10.1016/j.spl.2025.110516.

[7] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” 2021, *arXiv*. doi: 10.48550/ARXIV.2106.09685.

[8] Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[9] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Dec. 2023, pp. 9004–9017. [Online]. Available: https://openreview.net/forum?id=UMUeeSnsZm

[10] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” 2020, *arXiv*. doi: 10.48550/ARXIV.2005.11401.

[11] J. Pennington and P. Worah, “Nonlinear random matrix theory for deep learning,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124005, Dec. 2019, doi: 10.1088/1742-5468/ab3bc3.

[12] 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*, vol. 22, no. 165, pp. 1–73, 2021, [Online]. Available: http://jmlr.org/papers/v22/20-410.html

[13] T. N. Saada, A. Naderi, and J. Tanner, “Mind the Gap: a Spectral Analysis of Rank Collapse and Signal Propagation in Attention Layers,” 2024, *arXiv*. doi: 10.48550/ARXIV.2410.07799.

[14] A. Tamkin, D. Jurafsky, and N. Goodman, “Language Through a Prism: A Spectral Approach for Multiscale Language Representations,” Vancouver, Canada, 2020, pp. 1–15.

[15] K. Ethayarajh, “How Contextual are Contextualized Word Representations? Comparing the Geometry of

BERT, ELMo, and GPT-2 Embeddings," 2019, *arXiv*. doi: 10.48550/ARXIV.1909.00512.

[16] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," 2018, *arXiv*. doi: 10.48550/ARXIV.1802.05957.

[17] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep networks?," Montréal, Canada, 2018.

[18] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology (U.S.), Gaithersburg, MD, NIST AI 100-1, Jan. 2023. doi: 10.6028/NIST.AI.100-1.

[19] D. Voiculescu, "Limit laws for Random matrices and free products," *Inventiones mathematicae*, vol. 104, no. 1, pp. 201–220, Dec. 1991, doi: 10.1007/BF01245072.

[20] D. Voiculescu, "[No title found]," *International Mathematics Research Notices*, vol. 1998, no. 1, pp. 41–63, 1998, doi: 10.1155/S107379289800004X.

[21] "Caselaw Access Project 'Caselaw Dataset (Illinois).'" [Online]. Available: https://www.kaggle.com/datasets/harvardlil/caselaw-dataset-illinois

[22] Z. Chen *et al.*, "FinQA: A Dataset of Numerical Reasoning over Financial Data," 2021, *arXiv*. doi: 10.48550/ARXIV.2109.00122.

[23] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.

[24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia: PMLR 70, Jul. 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[25] K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," Montreal, Canada, 2018, pp. 1–20.

[26] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023, *arXiv*. doi: 10.48550/ARXIV.2307.09288.

[27] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[28] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[29] "The Business of Artificial Intelligence," *Harvard Business Review*, Jul. 18, 2017. [Online]. Available: https://hbr.org/2017/07/the-business-of-artificial-intelligence

Biographical notes

Kaushik Bar is a Gold Medalist MBA from IIM, Bangalore, and completed his B.E. in Electronics & Telecommunication Engineering from Jadavpur University, Kolkata. He holds US patents in data storage technology. He is the Co-Founder, CTO, and Chief Data Scientist at Inxite-Out Pvt Ltd, Kolkata. He has spent 20+ years in the software industry in firms like Intel, AMD, SanDisk etc. His areas of interest include Bayesian Optimization, Predictive Models, Generative AI, Agentic AI, and Embedded Systems.