Spectral Geometry of Attention: From Information Routing to Uncertainty

Giulio Viganò[*]
University of Milano-Bicocca
Milan, Italy
g.vigan36@campus.unimib.it

Simone Melzi
University of Milano-Bicocca
Milan, Italy

Maks Ovsjanikov
LIX, Ecole Polytechnique
IP Paris

Abstract

In this work, we study transformer attention through the lens of spectral geometry and operator theory. We view each attention head as a functional map between Hilbert spaces of functions on the token sequence and derive a *Token Difference Operator*, whose spectral structure controls how token-space information is routed to the output. We show that standard Euclidean spectra are structurally biased by sinks, conflating mass concentration with genuine routing capacity. By recasting token space in the intrinsic probability geometry induced by attention, the token difference spectrum disentangles sink effects from routing capacity and provides a spectral description of the dimensionality of the head output. This yields a unified framework for analyzing attention maps, explaining sinks, routing collapse, and output dimensionality within a single operator-theoretic framework. In practice, by grounding attention heuristics in spectral geometry, we develop a novel attention-based uncertainty estimator that complements probability-based scores, with the largest gains on long-context inputs.

1 Introduction

Attention maps are increasingly viewed as windows into the internal computation of transformer architectures [39]. Prior works have shown that attention carries useful signals about how transformers process information, studying it as probability distributions [39], graph operators [28, 7], transition matrices [15], or through row-wise statistics [40, 45]. While these perspectives have revealed important insights, they primarily focus on pairwise token interactions, or on models that do not describe the single-pass effect of attention, leaving comparatively unexplored the global structural properties of attention operators as a whole, especially in relation to the hidden representation dimensionality and the final output of the transformer.

In this work, we approach attention from this global operator perspective, framing each attention map A as a linear functional operator that transfers signals between geometric domains defined for the token sequence, where the geometry of each domain is the measure defining its inner product, and investigating its spectral properties. Inspired by functional maps [29] and shape difference operators [35, 21], we derive the *Token Difference Operator* (TDO) A^*A, a functional operator that characterizes the single-step action $V \mapsto AV$ induced by an attention head. Its spectrum reveals the multi-scale, global behavior of attention and relates, as we demonstrate, to a range of phenomena from attention sinks to model uncertainty.

The TDO departs from standard Euclidean analysis by assigning distinct geometric structures to the input and output domains of the attention map A. We demonstrate that the conventional practice of adopting a uniform Euclidean metric is biased by the inherent row-stochasticity of attention and the

[*]Work performed while at École Polytechnique

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

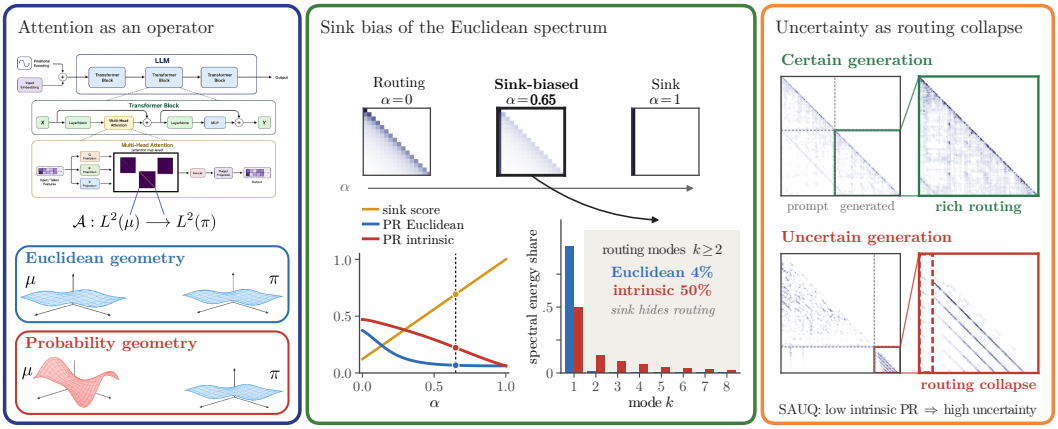


Figure 1: The spectral geometry of attention. *Left:* an attention head is a functional map $\mathcal{A} : L^2(\mu) \to L^2(\pi)$ between spaces of functions on the token sequence, and its properties depend on the measures placed on the two spaces: the Euclidean choice (blue) or an intrinsic probability geometry induced by attention itself (red). *Center:* a row-stochastic causal map interpolating (α) between a routing map and a sink map. The Euclidean spectrum is biased towards the sink; reweighting the input measure reduces the bias. *Right:* the same spectrum yields an uncertainty estimator, which reads uncertainty as a collapse of routing capacity in selected heads (red box in the uncertain map).

presence of attention sinks. In these biased settings, the mass assigned to initial tokens can saturate the operator spectrum, making it appear that there is significant information flow when no actual token-dependent routing is occurring. To rectify this, we induce an intrinsic probability geometry on the input domain, directly derived from the attention mechanism. With this intrinsic metric, the trivial constant mode corresponds to a transported mean, while the remaining centered modes describe the token-dependent component of the head output. This yields a mean–routing decomposition of AV that cleanly separates phenomena often conflated in the literature, including attention sinks [19, 4, 34] and dimensional collapse [12, 37].

Importantly, these structural quantities are not purely conceptual constructions: the TDO can be computed from attention maps, and its spectrum provides a strong signal in a range of probing tasks. We show that intrinsic spectral diagnostics are less dominated by sink score than raw attention or Euclidean spectral analysis, while aligning more closely with the dimensionality of the head output. Building on this, we introduce *Spectral Attention Uncertainty Quantification* (SAUQ), an unsupervised single-pass uncertainty estimator. Unlike probability-based estimators [46, 23, 14], which can fail when task-relevant distinctions are compressed before reaching the logits [3, 30, 41], SAUQ looks upstream at routing capacity and improves over closely related attention-based methods on summarization, in particular with long inputs.

We formalize these insights through the following theoretical and practical contributions, also summarized in Figure 1:

- **A global operator-theoretic view of attention.** We model attention maps as linear functional operators on token-space signals, shifting focus from token-level summaries to the spectral structure of the TDO.
- **A theoretical separation between sinks and dimensional collapse.** We prove that Euclidean spectra can be biased by sink tokens, and derive an intrinsic probability geometry in which the information dynamic is decomposed into a mean mode and a token-dependent routing subspace, proving that the nontrivial spectrum constrains the dimensionality of the head output.
- **A link to uncertainty:** We empirically show that the spectral signal of high routing heads can be leveraged to rank LLM generations by uncertainty.

2 Related Work

Attention Interpretability. A broad line of work studies attention as a mechanism for information processing and propagation in transformers. The classical view interprets attention weights as soft

similarity scores between queries and keys [39]. More recent approaches model attention as a discrete-time Markov chain [15], as a graph signal operator with an associated Laplacian [28, 7, 38], or through spectral properties such as QK eigenspectrum concentration and localization [2]. These methods reveal important structure, but they typically analyze attention as a probability matrix, graph, or repeated operator. Our work instead focuses on the single-step routing map $V \mapsto AV$ and studies its spectrum with respect to an intrinsic geometry induced by attention itself.

This contribution is more relevant in the presence of attention sinks and representation collapse. Attention sinks have been studied through embedding spectra [8], information propagation [4], and geometric analyses [34], and related phenomena have also been observed in vision transformers [11]. In parallel, hidden-state collapse and compression effects have been linked to oversquashing and attention sinks [3, 12], as well as to spiking behavior in feed-forward layers [37, 13]. We build on this line of work by showing that standard Euclidean spectral diagnostics can confuse sink concentration with routing capacity, and by introducing an intrinsic spectrum that separates these effects.

Uncertainty quantification in LLMs. Uncertainty estimation and hallucination detection for LLMs are often based on output probabilities, using quantities such as entropy, perplexity, likelihood, or sampling-based semantic consistency [46, 23, 14], including spectral signals computed on similarity graphs over multiple sampled responses [26] and tensor-network formulations of semantic uncertainty [42]. However, probability-based signals can be unreliable when task-relevant distinctions are compressed before reaching the logits [3, 30, 41]. This motivates uncertainty signals derived from internal model dynamics.

For this reason, recent methods leverage hidden representations of LLMs to quantify uncertainty [24, 9]. In particular, some works focus on attention heads, looking at attention values [36], previous-token relations [40], Laplacian eigenvalues of attention graphs [7], and topological divergence measures on attention graphs [5]. While closely related to these attention-based methods, our approach replaces scalar or graph-based summaries with the intrinsic spectrum of the TDO. Our solution grounds uncertainty estimation in the same geometry that controls routing capacity, yielding a single-pass estimator that captures spectral collapse in token-level information flow.

3 Theoretical Framework

We consider a decoder-only transformer with L layers and H attention heads per layer. The attention head h at layer l is computed as

$$A_{l,h} = \mathrm{softmax}\Big(Q_{l,h}K_{l,h}^{\top}/\sqrt{d} + C\Big), \tag{1}$$

and acts on the value matrix as $A_{l,h}V_{l,h}$, where C is the causal mask, $Q_{l,h}, K_{l,h}, V_{l,h} \in \mathbb{R}^{n,d}$. For each layer l, the full multi-head attention output is obtained by concatenating the heads and projecting the output $\mathrm{MHA}_l(\hat{X}_{l-1}) = \mathrm{concat}_h(A_{l,h}V_{l,h})\, W_l^O$. The complete transformer block notation is provided in the Supplementary material. Our study focuses on the matrix $A_{l,h}$ acting on the token space as a token-routing operator. We analyze one head at a time and omit the layer and head indices when no confusion arises. For a fixed input, each attention map $A \in \mathbb{R}^{n\times n}$, as an operator acting on V, is row-stochastic, causal, and linear in the columns of its input V. These properties correspond to:

$$A_{ij} \geq 0\ \forall i,j; \qquad \sum_{j=1}^{n} A_{ij} = 1\ \forall i; \qquad A_{ij} = 0\ \forall j > i. \tag{2}$$

3.1 Attention as a functional mapping and Token Difference Operator

Let $\mathcal{T} = \{1, \ldots, n\}$ denote the space of tokens described by their position. We define the weighted Hilbert space $L^2(\mu) = L^2(\mathcal{T}, \mu)$, as the space $\mathcal{T}$ equipped with the inner product:

$$\langle f, g\rangle_\mu = \sum_{j=1}^{n} \mu_j f_j g_j = f^{\top}\Omega_\mu g, \qquad f, g \in L^2(\mathcal{T}) = \mathbb{R}^n, \tag{3}$$

where μ is a positive measure on $\mathcal{T}$, which can be encoded as a vector of length n and $\Omega_\mu = \mathrm{diag}(\mu)$. Given another positive measure π on $\mathcal{T}$, we can interpret the attention matrix A as a linear map

$$\mathcal{A} : L^2(\mu) \longrightarrow L^2(\pi), \qquad \mathcal{A}(f) = Af. \tag{4}$$

Assuming $\mu_j > 0$ for all j, the operator $\mathcal{A}$ admits an adjoint $\mathcal{A}^* : L^2(\pi) \to L^2(\mu)$ defined by $\langle \mathcal{A}f, g\rangle_\pi = \langle f, \mathcal{A}^* g\rangle_\mu$, which can be computed with the following formula:

$$\mathcal{A}^* = \Omega_\mu^{-1} A^\top \Omega_\pi. \tag{5}$$

Having established these two operators and inspired by spectral geometry [21], we define the *Token Difference Operator* (TDO) as

$$D_{\mu,\pi} : L^2(\mu) \to L^2(\mu) \qquad D_{\mu,\pi} := \mathcal{A}^*\mathcal{A} = \Omega_\mu^{-1} A^\top \Omega_\pi A, \tag{6}$$

This operator is self-adjoint and positive semidefinite in $L^2(\mu)$, and is strictly related to the Rayleigh quotient $\mathcal{R}_{\mu,\pi}(f) = \|\mathcal{A}f\|_\pi^2 / \|f\|_\mu^2$, which measures how much the attention operator A stretches an input signal f by acting on it and transferring it between the two geometries. The properties of the TDO strongly depend on the measures μ, π, which we refer to as the geometry, and we will discuss this relation in the next sections.

We note that all spaces considered here are finite-dimensional: $L^2(\mu)$ is $\mathbb{R}^n$ equipped with a weighted inner product, and every result below can be written in matrix form. We adopt the operator language because it makes the construction discretization-agnostic, as for functional maps [29] and shape difference operators [35]; concurrent work [43] similarly treats attention as a map between function spaces.

3.1.1 Euclidean scenario: structural sink bias

Taking both μ and π as counting measures (i.e. $\mu_j = \pi_j = 1\ \forall j$) gives $D_{\mu,\pi} = A^\top A$, the standard Euclidean Gram matrix from which the standard singular values $\{\sigma_i(A)\}_{i=1}^n$ are computed. In this geometry, row-stochastic attention is generally not a contraction.

Proposition 1 (Euclidean Non-contractivity)**.** For A row-stochastic, $\sigma_1(A) \geq 1$, with equality if and only if A is doubly stochastic.

The constraint $A\mathbf{1} = \mathbf{1}$ fixes the constant function as a unit-gain direction for A, but in counting geometry the TDO acts on this mode as $A^\top A\mathbf{1} = A^\top \mathbf{1} = c, c_j = \sum_i A_{ij}$. Thus, the Euclidean Gram operator converts the constant mode into the column-sum profile. Under causal masking $A_{11} = 1$, hence $c_1 \geq 1$ and $\sum_j c_j = n$; early keys are also visible to more query positions than late keys. The leading Euclidean spectrum is therefore sensitive to accumulated sink mass. In the extreme causal sink $A = \mathbf{1}e_1^\top$, this becomes $A^\top A = ne_1e_1^\top$, so $\sigma_1(A)^2 = n$ and all spectral energy lies in the sink coordinate. The bias is not only a rank error, but a quadratic amplification of column-mass concentration. The proof of Proposition 1 and a Frobenius budget refining this picture appear in Supplementary materials A.3.

3.1.2 Probability geometry scenario: intrinsic non-expansion

The non-contractivity above reflects a mismatch between row-stochasticity and an unweighted geometry: the counting measure ignores the nonuniform column mass induced by causal attention. We resolve this by taking the output measure to be probabilistic and pulling it back to obtain the input measure:

$$\pi = \frac{1}{n}\mathbf{1}, \qquad \mu = A^\top \pi = \frac{1}{n}\sum_{i=1}^{n} (A_{i:})^\top. \tag{7}$$

Both are probability measures; μ is the average attention row over keys. This choice also restores the correct stationary geometry of the TDO. Since $\mu = A^\top \pi$, we have $D_{\mu,\pi}\mathbf{1} = \mathbf{1}, \mu^\top D_{\mu,\pi} = \mu^\top$. Thus, the pulled-back column-mass profile is not amplified as an unweighted Euclidean direction; it becomes the invariant measure with respect to which constants and centered fluctuations are defined.

Proposition 2 (Intrinsic non-expansion)**.** With the measures (7), $\mathcal{A} : L^2(\mu) \to L^2(\pi)$ satisfies $\|\mathcal{A}f\|_\pi \leq \|f\|_\mu$ for all $f \in L^2(\mu)$, with operator norm 1 and $\lambda_1(D_{\mu,\pi}) = 1$ attained by constant $\mathbf{1}$.

Sketch of proof. The bound follows from Jensen's inequality $(Af)_i^2 \leq \sum_j A_{ij} f_j^2$ averaged with uniform π; equality on $\mathbf{1}$ is trivial. Full proof in Supplementary materials A.4.

Corollary 1 (Centered contraction)**.** Let $\mathbf{1}^{\perp_\mu} = \{f \in L^2(\mu) : \mu^\top f = 0\}$ and order $1 = \lambda_1 \geq \lambda_2 \geq \cdots$. Then

$$\sup_{f \in \mathbf{1}^{\perp_\mu} \setminus \{0\}} \frac{\|\mathcal{A}f\|_\pi}{\|f\|_\mu} = \sqrt{\lambda_2(D_{\mu,\pi})} \leq 1. \tag{8}$$

The trivial unit-gain mode is the constant function. All non-trivial routing happens on its μ-orthogonal complement, where the contraction factor is governed by $\lambda_2(D_{\mu,\pi})$. We report the complete proof in the Supplementary materials A.5.

The unit singular mode is still present, but it now has a fixed interpretation: it is a constant function. This mode carries no token-dependent information, since it assigns the same value to every position and is preserved by any row-stochastic attention map. Therefore, the relevant question is not whether A preserves constants, but how much variation around the attention-weighted mean can survive one attention step.

3.2 From functions to representations

The preceding analysis concerns scalar functions $f \in L^2(\mathcal{T}, \mu)$. The value matrix $V \in \mathbb{R}^{n\times d}$ can be viewed columnwise as d such functions, hence, the intrinsic contraction result directly describes how an attention head routes hidden representations across token positions. We decompose V into its μ-mean and its centered component:

$$m := \mu^\top V \in \mathbb{R}^d, \qquad V_c := V - \mathbf{1}m, \qquad \mu^\top V_c = 0. \tag{9}$$

Here, m is the attention-weighted mean value vector, while V_c contains the token-dependent deviations around this mean.

Proposition 3 (Mean–routing decomposition)**.** Under the intrinsic measures (7), the head output decomposes as

$$AV = \mathbf{1}m + AV_c, \tag{10}$$

with $\langle \mathbf{1}m, AV_c\rangle_{\pi,F} = 0$ and the routed component bounded by

$$\|AV_c\|_{\pi,F} \le \sqrt{\lambda_2(D_{\mu,\pi})}\, \|V_c\|_{\mu,F}. \tag{11}$$

where we define $\|U\|^2_{\mu,F} = \mathrm{tr}(U^\top \Omega_\mu U)$ as the weighted Frobenius norm.

Sketch of proof. $V = \mathbf{1}m + V_c$ and $A\mathbf{1} = \mathbf{1}$ gives the split; orthogonality is trivial, and the norm bound applies Corollary 1 columnwise. Full proof in Supplementary materials A.6.

The choice of intrinsic measures aligns the barycenters of the input and output representations. Since $\pi^\top A = \mu^\top$, we have

$$\pi^\top AV = \mu^\top V = m. \tag{12}$$

Thus, attention maps the μ-weighted input mean to the π-weighted output mean. Thus, $\mathbf{1}m$ captures the position-independent transported mean, while AV_c captures token-dependent routing. The factor $\sqrt{\lambda_2(D_{\mu,\pi})}$ controls how much centered variation survives one attention step, and therefore governs the routed dimensionality, as shown in the following Corollary.

Corollary 2 (Spectral entropy bound)**.** Let $Z = \Omega_\pi^{1/2} AV$, $Z_c = \Omega_\pi^{1/2} AV_c$, and $\alpha = \|Z_c\|_F^2/(\|m\|_2^2 + \|Z_c\|_F^2)$. Define the spectral entropy $H(Y) = -\sum_k p_k \log p_k$ with $p_k = \sigma_k(Y)^2/\|Y\|_F^2$. Then, with h_2 the binary entropy,

$$\alpha H(Z_c) \le H(Z) \le h_2(\alpha) + \alpha H(Z_c), \qquad \|Z_c\|_F \le \sqrt{\lambda_2(D_{\mu,\pi})}\|V_c\|_{\mu,F}. \tag{13}$$

The spectrum of $D_{\mu,\pi}$ thus controls the dimensionality of the routed component in two complementary senses: as a norm constraint and as the spectral entropy of Z_c entering the mixture bound on $H(Z)$. Proof in Supplementary materials A.7.

3.3 Spectral metrics

The TDO spectrum describes how centered token-dependent information survives an attention step. We summarize it using the normalized participation ratio

$$\mathrm{PR}(D_{\mu,\pi}) := \frac{1}{n}\frac{\left(\sum_k \lambda_k\right)^2}{\sum_k \lambda_k^2} = \frac{1}{n}\frac{\mathrm{tr}(D_{\mu,\pi})^2}{\mathrm{tr}(D_{\mu,\pi}^2)}. \tag{14}$$

It ranges from $1/n$ for a single active mode to $\mathrm{rank}(D_{\mu,\pi})/n$, for a uniform nonzero spectrum, and is invariant to global rescaling. In practice, we compute it from the matrix $S_{\mu,\pi} =$

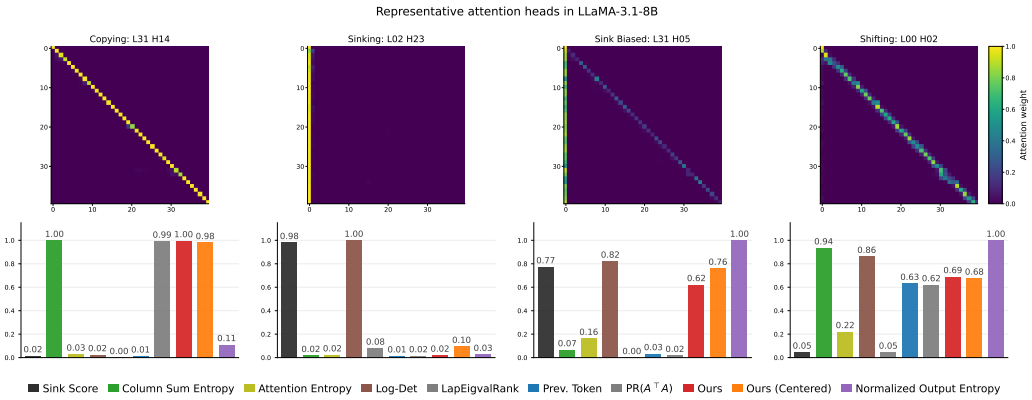


Figure 2: Representative attention heads and how different metrics classify their routing behavior. The intrinsic TDO score separates sink collapse from sink-biased routing, assigning high routing scores to copying, shifting, and sink-biased heads while distinguishing them from pure sink heads.

$\Omega_\mu^{-1/2} A^\top \Omega_\pi A \Omega_\mu^{-1/2}$, which is symmetric and similar to $D_{\mu,\pi}$. Large $\mathrm{PR}(D_{\mu,\pi})$ indicates that centered variation is routed through several active directions; small values indicate compression of V_c before the residual stream (Corollary 2). Since the constant mode carries no token-dependent information (Proposition 2), we also consider the centered normalized participation ratio, which excludes it,

$$\mathrm{PR}_{\mathrm{cent}}(D_{\mu,\pi}) = \frac{1}{n-1} \frac{\left(\sum_{k=2}^{n} \lambda_k\right)^2}{\sum_{k=2}^{n} \lambda_k^2} = \frac{1}{n-1} \frac{\left(\mathrm{tr}(S_{\mu,\pi}) - 1\right)^2}{\|S_{\mu,\pi}\|_F^2 - 1}, \tag{15}$$

where the second form uses $\lambda_1 = 1$. This is the statistic used in the uncertainty estimator of Section 5.

4 Sinks, Routing, and Output Dimensionality

We test two predictions of the theory: that the intrinsic TDO reduces sink bias, and that its centered spectrum tracks output dimensionality. We compare intrinsic spectral diagnostics ($\mathrm{PR}(D_{\mu,\pi})$ and its centered restriction) against (i) *raw attention statistics*, sink score, previous-token score [40], attention entropy [45]; and (ii) *counting-geometry spectral diagnostics*,log-determinant score [36], Laplacian diagonal entropy [7], and $\mathrm{PR}(A^\top A)$. All metrics are evaluated against both sink score and the output spectral entropy $H(AV)$ [3] across the full head population.

Representative heads Figure 2 compares four heads with clear routing properties that metrics should distinguish: a copying head, a sink head, a sink-biased head, and a shifting head. In the sink-biased head, substantial sink mass coexists with nontrivial routing, so counting-geometry and graph-based diagnostics confuse column-mass concentration with routing collapse, scoring it comparably to the sink head. The shifting head exposes the complementary failure: previous-token score [40] is tailored to this specific pattern and correctly flags it, but for the same reason misses the copying head, which routes information through a different position. In contrast, $\mathrm{PR}(D_{\mu,\pi})$ assigns high routing scores to the copying, shifting, and sink-biased heads, while separating them from sink collapse. Routing capacity does not by itself fix the output dimensionality: the copying head has a high routing score but low output dimensionality, because $H(AV)$ also depends on the properties of V.

Sink dependence and output dimensionality We first refine the relation between sink concentration and routing collapse. Under causal row-stochasticity, rank deficiency of A forces both phenomena jointly: a low-rank A must concentrate column mass on early tokens, in the rank-one limit, $A = \mathbf{1}e_1^\top$, and the TDO spectrum loses centered modes accordingly. The converse fails: a head with imbalanced column sums can still maintain a rich centered spectrum, so sink behavior does not imply representational collapse. Disentangling the two regimes, therefore, requires a diagnostic that does not amplify column mass as a Euclidean direction. The intrinsic TDO satisfies this: its leading mode is fixed at the constant function and routing is measured on the centered subspace

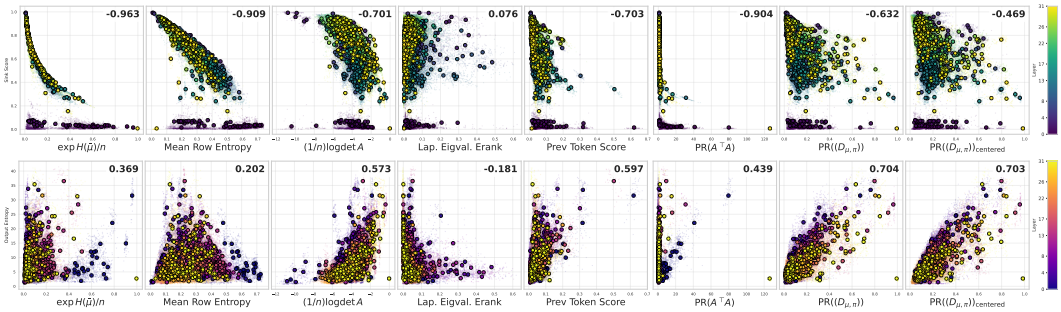


Figure 3: Scatterplots showing the relationships between attention diagnostics, sink score, and output dimensionality. Top: relation between each diagnostic and sink score. Bottom: relation between each diagnostic and the spectral entropy of the head output $H(AV)$. Points correspond to the attention heads of Llama-3.1-8B on 200 inputs, shown both per input and averaged across inputs (bold points). We report the Spearman correlation between the metrics.

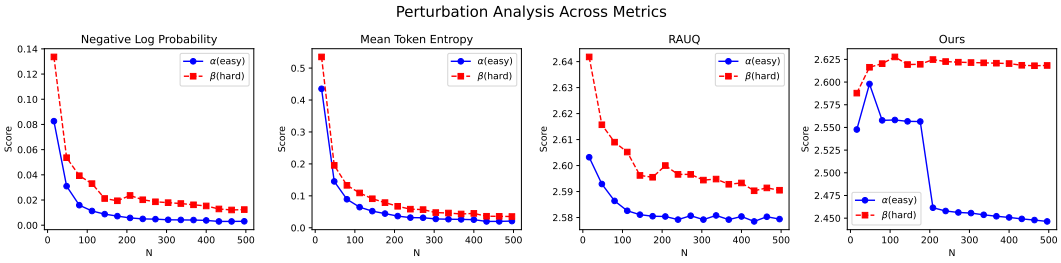


Figure 4: Perturbation sensitivity in a controlled copy task inspired by [41, 30]. The model receives copy: $b_1 b_2 \cdots b_n$ and must reproduce the bit string; we contrast the all-zero input 0^n (α, easy) with a perturbation $b^{(r)}$ carrying a single one at position r (β, hard). As n grows, perplexity and mean token entropy become nearly indistinguishable across the two families, while RAUQ [40] and Ours, computed from hidden-layer routing geometry, remain sensitive to the perturbation.

$\mathbf{1}^{\perp_\mu}$. Figure 3 confirms this empirically across heads. *Top:* column-sum entropy, attention entropy, previous-token score, log-determinant score, and $\mathrm{PR}(A^\top A)$ all show visible dependence on sink score, with Spearman correlations between -0.70 and -0.96. Diagnostics derived from $D_{\mu,\pi}$ are less sink-dependent: $\mathrm{PR}(D_{\mu,\pi})$ moderately so (-0.63), and its centered variant markedly so (-0.47). Laplacian diagonal entropy [7] shows the weakest sink dependence, plausibly because its degree normalization plays a role analogous to the pullback measure μ. *Bottom:* $\mathrm{PR}(D_{\mu,\pi})$ and its centered variant align more closely with $H(AV)$ (0.70) than raw attention statistics or counting-geometry ranks, supporting Corollary 2. The two variants are equally aligned on Llama-3.1-8B (0.704 vs. 0.703), and the centered one is slightly lower on Pythia and Qwen (Appendix B.1); its advantage is the lower sink dependence, not a closer alignment with output dimensionality. The baselines fail in complementary ways: raw entropies are dominated by sink structure, while log-determinant and Laplacian-based scores identify collapsed outputs but miss the full range of output entropies. The intrinsic TDO spectrum thus best captures the non-collapsed, token-dependent routing signal needed for the uncertainty estimator developed next.

5 Spectral Attention Uncertainty Quantification

Probability-based uncertainty estimators, such as perplexity, sequence likelihood, and token entropy, are computed at the unembedding output, after the residual stream has been mapped to logits. When task-relevant distinctions are compressed during representation propagation, two inputs that differ in important ways induce nearly indistinguishable logits, and any output-probability estimator inherits this collapse [3, 30, 41], especially in long-range or non-local tasks. Figure 4 makes this concrete: on a controlled copy task, perplexity and token entropy become indistinguishable across two input families that still differ in a task-relevant position, while attention-based signals derived from hidden-layer routing geometry remain sensitive.

The framework of Section 3 defines what the signal measures: the TDO spectrum quantifies the effective dimensionality of the centered, token-dependent component routed by an attention head. We hypothesize that spectral collapse during generation signals a loss of routing capacity that is not yet visible in output probabilities, and we test this hypothesis empirically. To this end, we define *Spectral Attention Uncertainty Quantification* (SAUQ), a single-pass estimator that turns intrinsic attention spectra into a confidence score for selective prediction.

5.1 Method

Let $x = (x_1, \ldots, x_m)$ be the prompt and $y = (y_1, \ldots, y_T)$ the generated continuation. At the decoding step t, the model attends over the prefix of length $n_t = m + t$, with attention map $A_t^{(l,h)} \in \mathbb{R}^{n_t \times n_t}$ for layer l and head h. We denote the generation-only submatrix of $A_t^{(l,h)}$ as $B_t^{(l,h)} = A_t^{(l,h)}[G_t, G_t]$, with $G_t = [m+1, \ldots, m+t]$ the set of generated tokens.

Head selection. We focus on the middle third of transformer layers, $\mathcal{L}_{\text{mid}}$, as in RAUQ [40] (layers 10-22 for Llama-3.1-8B). Within each such layer, we first select a routing head using a lightweight spectral proxy computed only on the generated-token attention block. Let $B_T^{(l,h)}$ be the generated-token square block at the final decoding step. We select

$$h^*(l) = \arg\max_h \mathrm{PR}\Big(B_T^{(l,h)} B_T^{(l,h)\top}\Big). \tag{16}$$

This proxy is a cheap empirical spectral screen: it scores all H heads of a layer with a single batched product on $T \times T$ blocks, whereas the intrinsic signal below is computed only for the selected head.

Spectral signal. For each selected head and at each generation step, we equip the full prefix with the intrinsic measures of (7), $\pi_t = \frac{1}{n_t}\mathbf{1}$ and $\mu_t = A_t^{(l,h^*)\top}\pi_t$, and compute the centered normalized participation ratio of the TDO (15)

$$s_t^{(l)} := \mathrm{PR}_{\text{cent}}\Big(D_{\mu_t,\pi_t}^{(l,h^*)}\Big). \tag{17}$$

We use the centered variant because it discards the trivial constant mode and is less sink-dependent than the full PR 4, which performs on par with it (Table 1). The signal only requires $\mathrm{tr}(S_{\mu_t,\pi_t}) - 1$ and $\|S_{\mu_t,\pi_t}\|_F^2 - 1$, so no eigendecomposition is needed.

Confidence recurrence. Then, we integrate our signal into the recurrence structure of RAUQ [40]. Let $p_t = P(y_t \mid y_{<t}, x)$ be the model probability assigned to the generated token at step t. SAUQ defines a layer-wise confidence recursively by

$$c_t^{(l)} = \begin{cases} p_1, & t = 1, \\ \alpha\, p_t + (1-\alpha)\, s_t^{(l)} c_{t-1}^{(l)}, & t > 1, \end{cases} \tag{18}$$

where $\alpha \in [0,1]$ trades off local token probability against confidence propagated through the spectral routing structure; in the following experiments, we select $\alpha = 0.2$ following [40]. The recurrence, α, and the layer range are taken unchanged from RAUQ: they are not dictated by our theory, and keeping them fixed means that SAUQ differs from RAUQ only in the head-selection rule and in the routing signal.

Under this recurrence, high spectral participation preserves prior confidence, reflecting the presence of multiple active routing directions. By contrast, spectral collapse attenuates propagated confidence, since the head routes information through fewer effective degrees of freedom. For each selected layer, we convert the confidence trajectory into an uncertainty score by averaging log-confidence over the generated sequence, $U_{\text{SAUQ}}^{(l)} = 1 - \frac{1}{T}\sum_{t=1}^{T} \log c_t^{(l)}$. The final sequence-level score is the maximum uncertainty over the middle-third layers, $U_{\text{SAUQ}} = \max_{l \in \mathcal{L}_{\text{mid}}} U_{\text{SAUQ}}^{(l)}$. We use the maximum because SAUQ is a rejection score: a routing collapse in any monitored head should raise uncertainty rather than be averaged away. Since our analysis is per head and precedes residual mixing and MLP updates, such a collapse need not imply information loss for the whole model. Thus, SAUQ preserves the recurrent structure of RAUQ, but replaces its scalar previous-token score with the effective spectral dimension of routing heads from Section 3. SAUQ thus yields per-step

Estimator	Category	QA					Summarization		
		TriviaQA	TruthfulQA	MMLU	CoQA	Mean	SAMSum	XSum	Mean
AttentionScore [36]	Single pass	0.0796	-0.0505	0.2078	0.0933	0.0826	-0.3727	-0.9983	-0.6855
MaximumSequenceProbability	Single pass	0.2518	0.3211	0.5501	0.2878	0.3527	-0.1363	-0.386	-0.2612
MeanTokenEntropy	Single pass	0.5661	**0.4199**	0.2437	0.2568	0.3716	0.4999	0.3928	0.4464
Perplexity	Single pass	0.1088	0.2863	0.5158	0.2737	0.2962	0.5071	0.4053	0.4562
CCP[16]	Single pass	0.1029	0.159	0.5567	0.2602	0.2697	-0.1458	-0.3018	-0.2238
NLI-DegMat [26]	Sample-based	0.5789	0.2061	0.4381	0.3959	0.4048	0.3888	0.3236	0.3562
NLI-Eccentricity [26]	Sample-based	0.5641	0.2007	0.4469	0.3805	0.3981	0.344	0.0351	0.1896
NLI-EigValLaplacian [26]	Sample-based	**0.6034**	0.1962	0.4681	**0.4147**	**0.4206**	0.3864	0.3845	0.3855
EigenScore [9]	Sample-based	0.3611	0.0524	0.4118	0.3467	0.293	0.0719	-0.2122	-0.0702
LUQ [46]	Sample-based	0.5878	0.182	0.4392	0.3588	0.392	0.29	0.3585	0.3243
LexicalSimilarity_rougeL [17]	Sample-based	0.4788	0.1599	0.4127	0.3636	0.3538	0.2483	0.1981	0.2232
SAR [14]	Sample-based	0.5733	0.1786	0.4103	0.3275	0.3724	0.2936	0.3069	0.3003
SemanticDensity[31]	Sample-based	0.5592	0.0755	0.1582	0.1102	0.2258	0.0752	0.1138	0.0945
SemanticEntropy[23]	Sample-based	0.2005	0.2113	0.4701	0.3662	0.312	0.2525	-0.1886	0.032
RAUQ [40]	Single pass	0.2631	0.3369	**0.5596**	0.311	0.3677	0.5269	0.3008	0.4139
LapEigvalRAUQ [7]	Single pass	0.1190	0.3296	0.5578	0.3064	0.3282	0.5285	0.2907	0.4096
LogDetRAUQ [36]	Single pass	0.1140	0.3313	**0.5596**	0.3087	0.3284	0.5288	0.3048	0.4168
SAUQ (gen. only)	Single pass	0.5013	0.3056	0.5375	0.311	0.4139	0.4948	0.2578	0.3763
SAUQ (full PR)	Single pass	0.2914	0.3767	0.5588	0.3055	0.3831	0.5924	**0.4214**	0.5069
SAUQ (Ours)	Single pass	0.2704	0.3876	0.5574	0.3046	0.3800	**0.6028**	0.4163	**0.5096**

Table 1: Comparison against unsupervised methods. The best result in each column is shown in bold, and the second best is underlined.

($c_t^{(l)}$) and per-layer ($U_{\mathrm{SAUQ}}^{(l)}$) scores, but it does not localize uncertainty within the spectrum: the participation ratio does not identify which eigenmodes collapse. Since the signal measures the routing of the model's own computation rather than the dispersion of its output distribution, we view it as epistemic-leaning; it does not provide a clean separation between epistemic and aleatoric uncertainty.

Generation-only variant. The full D_{μ_t,π_t} mixes prompt and generated tokens. To isolate the uncertainty accumulated during the generation, we also consider a variant in which the signal is the same Euclidean proxy used for head selection, evaluated at each step on the generated block, $s_t^{(l),\mathrm{gen}} := \mathrm{PR}\big(B_t^{(l,h^*)} B_t^{(l,h^*)\top}\big)$, with selection, recurrence and aggregation unchanged. Since the prompt columns are removed and B_t is not renormalized, B_t is sub-stochastic and the theory do not apply. In B.6, we analyze the effect of having different selection-signal relations in SAUQ.

5.2 Evaluation

We evaluate SAUQ using Llama-3.1-8B [1] and Qwen2.5-7B [32] (Appendix B.3). We use the LM Polygraph benchmark [16] to ensure easy reproducibility, following the details provided in the experimental setting of [40]. We evaluate our method on question-answering task on TriviaQA [22], TruthfulQA [25], MMLU [20], CoQA [33], and on the summarization task on SAMSum [18], and XSum [27]. We report the Normalized Prediction Rejection Ratio (PRR; higher is better). PRR evaluates the ranking induced by a score, and SAUQ is a ranking score. Indeed, on QA, we evaluate single-pass selective prediction, i.e., how well one score ranks the greedy answers. Table 1 compares SAUQ performances against single-step probability-based methods (Perplexity, MaxSequenceProbability, MeanTokenEntropy, CCP), sample-based (10 samples) methods (NLI-based spectral measures, EigenScore, LUQ, LexicalSimilarity, SAR, Semantic Density, SemanticEntropy), and attention-based baselines (AttentionScore, RAUQ, LapEigvalRAUQ, LogDetRAUQ). The last two are variants of RAUQ in which the previous-token statistic, used both for head selection and as the signal, is replaced by LapEigval [7] and the log-determinant AttentionScore [36], respectively.

In Table 1, we report the mean PRR values for Llama-3.1-8B; We refer to the Appendix for additional results. On QA datasets, SAUQ is competitive with strong unsupervised estimators, but it is not the best method on average; several sample-based baselines obtain stronger mean QA performance. In contrast, the gains are clearest on summarization, where SAUQ and its full-PR variant obtain the two highest scores on both SAMSum and XSum and the strongest summarization means among all methods considered. Part of this gain is a length effect, that we analyze in Appendix B.7. The residual part grows with input length, which is consistent with our motivation that the upstream routing spectrum retains distinctions across long contexts. We therefore attribute to long-range routing only this residual effect. We note that SAUQ is a complementary single-pass signal to both probability based and attention based signals.

SAUQ, RAUQ, LapEigvalRAUQ, and LogDetRAUQ share the recurrence, layer range, and aggregation, but each selects heads with its own attention statistic. Their comparison therefore reflects the head-selection rule and the routing signal jointly, not the signal alone; Appendix B.6 varies the two independently. The ranking is in line with the metric-quality findings of Section 4, where the LapEigval and LogDet signals track output dimensionality less closely than the intrinsic TDO.

Choice of measure The framework treats the input measures μ, π as an explicit modeling axis rather than a fixed convention. In the context of uncertainty, we tested alternative approaches to uniform and intrinsic measures, including masses based on token positions, frequency of tokens (Zipf's Law), and unembedding matrix decomposition. Definitions of all measures and an example of their distribution are given in Appendix B.8. Only the intrinsic choice $\mu = A^\top \pi$ satisfies Proposition 2; for the others the constant function is not an eigenvector of $D_{\mu,\pi}$. On short-range QA, several measures perform comparably, with positional weighting marginally best. On summarization, the attention-induced measure outperforms the alternatives, while counting geometry lags. Interestingly, the performances of frequency- and unembedding-based estimators suggest that the 'semantic' of the tokens is less relevant in this scenario with respect to positional bias. Extending such analysis forward offers an intriguing future direction in which our approach can be exploited for a broader understanding of LLMs' computations.

Measure	**QA**	**Summ.**
Uniform	0.3660	0.4204
Positional	**0.4027**	<u>0.4771</u>
Zipf-Law	0.3718	0.4223
Unembedding	0.3755	0.4202
Intrinsic	<u>0.3831</u>	**0.5069**

Table 2: Mean PRR of SAUQ under different measures (Llama-3.1-8B), computed with the full PR.

6 Conclusion

We introduced a spectral-geometric view of attention. Euclidean spectral diagnostics can be biased by sink behavior, conflating column-mass concentration with genuine routing; the intrinsic probability geometry instead yields a token-difference spectrum that separates the trivial constant mode from the centered modes controlling routing capacity. Empirically, the TDO better disentangles sink score from head-output dimensionality, and supports uncertainty estimation. Building on it, SAUQ replaces scalar attention probes with a spectral measure of routing capacity and improves over closely related attention-based baselines on summarization. The code is available at `https://github.com/gviga/spectral-geometry-attention`.

Limitations. Our analysis is per-head, per-input, and single-step, characterizing the routing map before residual mixing, MLP updates, and cross-head output projection; extending it to the full dynamics of representation evolution is a natural next step. The head-selection rule of SAUQ is empirical and not derived from the theory, and the gains are model dependent (Appendix B.3). Practically, the full TDO spectrum is more expensive than scalar attention statistics; SAUQ avoids the eigendecomposition by using only the participation ratio, but forming $S_{\mu,\pi}$ still costs $\mathcal{O}(n_t^3)$ per selected head and step.

Future work. Natural extensions include recovery-aware aggregation that accounts for routing restored in later layers, using the TDO spectrum to guide model editing, monitoring routing collapse during training, and applying the framework to attention in non-text domains.

Broader Impact. More faithful diagnostics of attention routing may improve interpretability, uncertainty estimation, and reliability assessment in transformer models, especially for deployment of safe and reliable AI systems.

Acknowledgments and Disclosure of Funding

This work was supported by the ERC Consolidator Grant 101087347 (VEGA), and by gifts from Ansys and Adobe Research. The MUR supported this work with the grant "Dipartimenti di Eccellenza 2023-2027" REGAINS of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca.

References

[1] AI@Meta. Llama 3 model card. 2024.

[2] Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates, 2024.

[3] Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G.M. Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc.

[4] Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token?, 2025.

[5] Alexandra Bazarova, Aleksandr Yugay, Andrey Shulga, Alina Ermilova, Andrei Volodichev, Konstantin Polev, Julia Belikova, Rauf Parchiev, Dmitry Simakov, Maxim Savchenko, Andrey Savchenko, Serguei Barannikov, and Alexey Zaytsev. Hallucination detection in llms with topological divergence on attention graphs, 2025.

[6] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

[7] Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Hallucination detection in LLMs using spectral features of attention maps. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24354–24385, Suzhou, China, November 2025. Association for Computational Linguistics.

[8] Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4792–4808, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection, 2024.

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

[11] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024.

[12] Enrique Queipo de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin, 2026.

[13] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023.

[14] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models, 2024.

[15] Yotam Erel, Olaf Dünkel, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Amit H. Bermano. Attention (as discrete-time markov) chains, 2025.

[16] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. LM-polygraph: Uncertainty estimation for language models. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 446–461, Singapore, December 2023. Association for Computational Linguistics.

[17] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555, 2020.

[18] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, page 70–79. Association for Computational Linguistics, 2019.

[19] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view, 2025.

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[21] R. Huang, F. Chazal, and M. Ovsjanikov. On the stability of functional maps and shape difference operators. *Computer Graphics Forum*, 37(1):145–158, 2018.

[22] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017.

[23] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023.

[24] Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Velickovic. How do llms compute verbal confidence, 2026.

[25] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[26] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024.

[27] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, 2018.

[28] Valentin Noël. A graph signal processing framework for hallucination detection in large language models, 2025.

[29] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Trans. Graph.*, 31(4), July 2012.

[30] Hector Pasten, Felipe Urrutia, Hector Jimenez, Cristian B. Calderon, Cristóbal Rojas, and Alexander Kozachinskiy. Continuity and isolation lead to doubts or dilemmas in large language models, 2025.

[31] Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space, 2024.

[32] Qwen Team. Qwen2.5 technical report, 2025.

[33] Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge, 2019.

[34] Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink, 2025.

[35] Raif M. Rustamov, Maks Ovsjanikov, Omri Azencot, Mirela Ben-Chen, Frédéric Chazal, and Leonidas Guibas. Map-based exploration of intrinsic shape differences and variability. *ACM Trans. Graph.*, 32(4), July 2013.

[36] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc., 2024.

[37] Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks, 2026.

[38] Huiyuan Tian, Bonan Xu, Shijian Li, and Gang Pan. Spectralkd: A unified framework for interpreting and distilling vision transformers via spectral analysis, 2025.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[40] Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms, 2025.

[41] Petar Veličković, Federico Barbero, Christos Perivolaropoulos, Simon Osindero, and Razvan Pascanu. Perplexity cannot always tell right from wrong, 2026.

[42] Pragatheeswaran Vipulanandan, Kamal Premaratne, and Dilip Sarkar. Semantic uncertainty quantification of hallucinations in llms: A quantum tensor network based method, 2026.

[43] Jiefang Xiao, Maolin Gao, Simon Weber, Guandao Yang, and Daniel Cremers. Functional attention: From pairwise affinities to functional correspondences, 2026.

[44] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[45] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse, 2023.

[46] Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. Luq: Long-text uncertainty quantification for llms, 2024.

Appendix for "Spectral Geometry of Attention: From Information Routing to Uncertainty".

A Proofs for the Theoretical Section

A.1 Full Notation

We consider a decoder-only transformer with L layers and H attention heads per layer, processing a sequence of n tokens. At layer l, the input is $X_{l-1} \in \mathbb{R}^{n\times d}$, whose i-th row x_i is the hidden representation of token i. We adopt the pre-norm convention as in [3]:

$$\hat{X}_{l-1} = \mathrm{LN}(X_{l-1}), M_l = X_{l-1} + \mathrm{MHA}_l(\hat{X}_{l-1}), \quad X_l = M_l + \mathrm{MLP}_l(\mathrm{LN}(M_l)). \tag{19}$$

For each head h, queries, keys, and values are $Q_{l,h} = \hat{X}_{l-1}W^Q_{l,h}$, $K_{l,h} = \hat{X}_{l-1}W^K_{l,h}$, $V_{l,h} = \hat{X}_{l-1}W^V_{l,h}$. The attention matrix is

$$A_{l,h} = \mathrm{softmax}\Big(Q_{l,h}K_{l,h}^\top/\sqrt{d} + C\Big), \tag{20}$$

where C is the causal mask, with $C_{ij} = 0$ for $j \le i$ and $C_{ij} = -\infty$ otherwise. The full multi-head attention output is obtained by concatenating the head and projecting outputs $\mathrm{MHA}_l(\hat{X}_{l-1}) = \mathrm{concat}_h(A_{l,h}V_{l,h})\, W^O_l$.

In this section we provide the technical details for the results stated in Section 3.

A.2 TDO is self-adjoint

We recall $\mathcal{A} : L^2(\mu) \longrightarrow L^2(\pi), \quad \mathcal{A}f = Af.$. Its adjoint is defined as the operator A^* such that $\langle \mathcal{A}f, g\rangle_\pi = \langle f, \mathcal{A}^*g\rangle_\mu$. We compute it explicitly

$$\langle \mathcal{A}f, g\rangle_\pi = (Af)^\top \Omega_\pi g = f^\top A^\top \Omega_\pi g = f^\top \Omega_\mu\big(\Omega_\mu^{-1}A^\top\Omega_\pi g\big) = \langle f, \Omega_\mu^{-1}A^\top\Omega_\pi g\rangle_\mu. \tag{21}$$

Hence, $\mathcal{A}^* = \Omega_\mu^{-1}A^\top\Omega_\pi$. The token-difference operator $D_{\mu,\pi} := \mathcal{A}^*\mathcal{A} = \Omega_\mu^{-1}A^\top\Omega_\pi A.$ is self-adjoint in $L^2(\mu)$. Indeed, for any $f, g \in L^2(\mu)$,

$$\langle D_{\mu,\pi}f, g\rangle_\mu = \langle \mathcal{A}^*\mathcal{A}f, g\rangle_\mu = \langle \mathcal{A}f, \mathcal{A}g\rangle_\pi = \langle f, \mathcal{A}^*\mathcal{A}g\rangle_\mu = \langle f, D_{\mu,\pi}g\rangle_\mu. \tag{22}$$

It is also positive semidefinite:

$$\langle D_{\mu,\pi}f, f\rangle_\mu = \langle \mathcal{A}f, \mathcal{A}f\rangle_\pi = \|Af\|^2_\pi \ge 0. \tag{23}$$

The associated Rayleigh quotient is

$$\mathcal{R}_{\mu,\pi}(f) = \frac{\|Af\|^2_\pi}{\|f\|^2_\mu} = \frac{f^\top A^\top\Omega_\pi Af}{f^\top\Omega_\mu f} = \frac{\langle D_{\mu,\pi}f, f\rangle_\mu}{\langle f, f\rangle_\mu}. \tag{24}$$

Thus the eigenvalues of $D_{\mu,\pi}$ describe the squared gains of attention under the input geometry μ and the output geometry π.

Although $D_{\mu,\pi}$ is not generally symmetric in the Euclidean inner product, it is similar to the symmetric positive semidefinite matrix

$$S_{\mu,\pi} := \Omega_\mu^{-1/2}A^\top\Omega_\pi A\Omega_\mu^{-1/2}. \tag{25}$$

Indeed,

$$D_{\mu,\pi} = \Omega_\mu^{-1/2}S_{\mu,\pi}\Omega_\mu^{1/2}. \tag{26}$$

Therefore $D_{\mu,\pi}$ and $S_{\mu,\pi}$ have the same eigenvalues. In practice, we can compute spectral quantities using $S_{\mu,\pi}$.

A.3 Euclidean case: proof of Proposition 1

When both μ and π are counting measures, $\Omega_\mu = \Omega_\pi = I$ and $D_{\mu,\pi} = A^\top A$. Thus, the eigenvalues of $D_{\mu,\pi}$ are the squared singular values of A.

Proof of Proposition 1. Because A is row-stochastic, $A\mathbf{1} = \mathbf{1}$. Therefore $\|A\mathbf{1}\|_2 = \|\mathbf{1}\|_2$. By the definition of the spectral norm,

$$\sigma_1(A) = \|A\|_2 = \sup_{f \neq 0} \frac{\|Af\|_2}{\|f\|_2} \geq \frac{\|A\mathbf{1}\|_2}{\|\mathbf{1}\|_2} = 1. \tag{27}$$

Now suppose that $\sigma_1(A) = 1$. Then $\|A\|_2 = 1$, and the vector $\mathbf{1}$ attains the operator norm. Equivalently, the Rayleigh quotient of $A^\top A$ at $\mathbf{1}$ equals its largest eigenvalue:

$$\frac{\mathbf{1}^\top A^\top A \mathbf{1}}{\mathbf{1}^\top \mathbf{1}} = 1 = \lambda_{\max}(A^\top A). \tag{28}$$

Since $A^\top A$ is symmetric positive semidefinite, equality in the Rayleigh principle implies that $\mathbf{1}$ is an eigenvector of $A^\top A$ with eigenvalue 1: $A^\top A\mathbf{1} = \mathbf{1}$. Using $A\mathbf{1} = \mathbf{1}$, this becomes$A^\top \mathbf{1} = \mathbf{1}$. Hence every column of A also sums to one, so A is doubly stochastic.

Conversely, suppose that A is doubly stochastic. Since A is nonnegative and has row sums and column sums equal to one,

$$\|A\|_\infty = 1, \qquad \|A\|_1 = 1. \tag{29}$$

The standard norm interpolation bound gives

$$\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty} = 1. \tag{30}$$

Together with the lower bound $\|A\|_2 \geq 1$, this gives $\sigma_1(A) = \|A\|_2 = 1$. □

Frobenius budget. The total squared singular mass is

$$\sum_k \sigma_k(A)^2 = \|A\|_F^2 = \sum_{i,j} A_{ij}^2. \tag{31}$$

Since each row of A is a probability vector, $0 \leq A_{ij} \leq 1$, and therefore $A_{ij}^2 \leq A_{ij}$. Hence

$$\|A\|_F^2 = \sum_{i,j} A_{ij}^2 \leq \sum_{i,j} A_{ij} = n. \tag{32}$$

Consequently,

$$\sum_{k \geq 2} \sigma_k(A)^2 = \|A\|_F^2 - \sigma_1(A)^2 \leq n - \sigma_1(A)^2. \tag{33}$$

Thus, when the leading singular value is large, a large fraction of the available Frobenius budget is concentrated in the leading mode.

There is also a lower bound. For any probability vector p supported on s coordinates, $\sum_{j=1}^s p_j^2 \geq \frac{1}{s}$, by Cauchy's inequality. Without causal constraints, every row may use all n positions, so $\|A\|_F^2 \geq 1$, with equality at the fully uniform matrix. Under causal masking, row i is supported on at most i positions. Therefore $\sum_j A_{ij}^2 \geq \frac{1}{i}$, and hence $\|A\|_F^2 \geq \sum_{i=1}^n \frac{1}{i}$.

A.4 Intrinsic probability geometry: proof of Proposition 2

We now use the intrinsic measures

$$\pi = \frac{1}{n}\mathbf{1}, \qquad \mu = A^\top \pi. \tag{34}$$

Proof of Proposition 2. For each query position i, the row $A_{i:}$ is a probability vector. Therefore, by Jensen's inequality applied to the convex function $x \mapsto x^2$,

$$(Af)_i^2 = \left(\sum_j A_{ij} f_j \right)^2 \leq \sum_j A_{ij} f_j^2. \tag{35}$$

Averaging this inequality over i with the uniform measure $\pi_i = 1/n$ gives

$$\|Af\|_\pi^2 = \frac{1}{n}\sum_{i=1}^n (Af)_i^2 \le \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n A_{ij} f_j^2 = \sum_{j=1}^n \left(\frac{1}{n}\sum_{i=1}^n A_{ij}\right) f_j^2 = \sum_{j=1}^n \mu_j f_j^2 = \|f\|_\mu^2. \quad (36)$$

Hence $\|\mathcal{A}\|_{L^2(\mu)\to L^2(\pi)} \le 1$. Since A is row-stochastic $A\mathbf{1} = \mathbf{1}$. Because both μ and π are probability measures their respective norm is 1. Therefore $\|\mathcal{A}\mathbf{1}\|_\pi = \|\mathbf{1}\|_\pi = \|\mathbf{1}\|_\mu$. So the operator norm is exactly one.

Finally, the eigenvalues of $D_{\mu,\pi}$ are squared singular gains of $\mathcal{A} : L^2(\mu) \to L^2(\pi)$. The previous inequality shows that all eigenvalues are at most one, and the constant function attains gain one. Hence $\lambda_1(D_{\mu,\pi}) = 1, \qquad D_{\mu,\pi}\mathbf{1} = \mathbf{1}$. □

The last identity can also be checked directly:

$$D_{\mu,\pi}\mathbf{1} = \Omega_\mu^{-1} A^\top \Omega_\pi A\mathbf{1} = \Omega_\mu^{-1} A^\top \Omega_\pi \mathbf{1} = \Omega_\mu^{-1} A^\top \pi = \Omega_\mu^{-1}\mu = \mathbf{1}. \quad (37)$$

Moreover,

$$\mu^\top D_{\mu,\pi} = \mu^\top \Omega_\mu^{-1} A^\top \Omega_\pi A = \mathbf{1}^\top A^\top \Omega_\pi A = (A\mathbf{1})^\top \Omega_\pi A = \mathbf{1}^\top \Omega_\pi A = \pi^\top A = \mu^\top. \quad (38)$$

Thus $\mathbf{1}$ is a right eigenvector and μ is the corresponding invariant left measure.

A.5 Proof of Corollary 1

Define the μ-centered subspace

$$\mathbf{1}^{\perp_\mu} = \{f \in L^2(\mu) : \mu^\top f = 0\}. \quad (39)$$

The identity $\pi^\top A = \mu^\top$ implies that attention maps centered functions to centered functions. Indeed, if $f \in \mathbf{1}^{\perp_\mu}$, then

$$\pi^\top Af = \mu^\top f = 0. \quad (40)$$

Thus Af is π-centered.

Next, since $D_{\mu,\pi}$ is self-adjoint in $L^2(\mu)$ and $D_{\mu,\pi}\mathbf{1} = \mathbf{1}$, the orthogonal complement $\mathbf{1}^{\perp_\mu}$ is invariant under $D_{\mu,\pi}$. To see this directly, take any $f \in \mathbf{1}^{\perp_\mu}$. Then

$$\langle D_{\mu,\pi} f, \mathbf{1}\rangle_\mu = \langle f, D_{\mu,\pi}\mathbf{1}\rangle_\mu = \langle f, \mathbf{1}\rangle_\mu = \mu^\top f = 0. \quad (41)$$

Hence $D_{\mu,\pi} f \in \mathbf{1}^{\perp_\mu}$.

For $f \in \mathbf{1}^{\perp_\mu}$, the squared gain of attention is

$$\frac{\|Af\|_\pi^2}{\|f\|_\mu^2} = \frac{\langle D_{\mu,\pi} f, f\rangle_\mu}{\langle f, f\rangle_\mu}. \quad (42)$$

By the Rayleigh-Ritz variational principle applied to the restriction of $D_{\mu,\pi}$ to $\mathbf{1}^{\perp_\mu}$,

$$\sup_{f\in\mathbf{1}^{\perp_\mu}\setminus\{0\}} \frac{\|Af\|_\pi^2}{\|f\|_\mu^2} = \lambda_{\max}\left(D_{\mu,\pi}\big|_{\mathbf{1}^{\perp_\mu}}\right). \quad (43)$$

With the eigenvalues ordered as

$$1 = \lambda_1 \ge \lambda_2 \ge \cdots \ge 0, \quad (44)$$

and with one copy of the eigenvalue $\lambda_1 = 1$ corresponding to the constant function, this restricted maximum is λ_2. Therefore

$$\sup_{f\in\mathbf{1}^{\perp_\mu}\setminus\{0\}} \frac{\|Af\|_\pi}{\|f\|_\mu} = \sqrt{\lambda_2(D_{\mu,\pi})} \le 1. \quad (45)$$

If the unit eigenvalue has multiplicity larger than one, then there exist nonconstant centered modes with unit gain, and the same formula gives $\lambda_2(D_{\mu,\pi}) = 1$.

A.6 Proof of Proposition 3

For matrix-valued token signals $U, W \in \mathbb{R}^{n\times d}$, define

$$\langle U, W\rangle_{\mu,F} = \operatorname{tr}(U^\top \Omega_\mu W), \qquad \|U\|^2_{\mu,F} = \operatorname{tr}(U^\top \Omega_\mu U). \tag{46}$$

The analogous π-weighted quantities are defined by replacing Ω_μ with Ω_π.

Given a value matrix $V \in \mathbb{R}^{n\times d}$, define its μ-mean and centered component by

$$m := \mu^\top V \in \mathbb{R}^d, \qquad V_c := V - \mathbf{1}m. \tag{47}$$

Then

$$\mu^\top V_c = \mu^\top V - \mu^\top \mathbf{1}m = m - (\mathbf{1}^\top \mu)m = m - m = 0, \tag{48}$$

because μ is a probability measure. Thus every feature column of V_c lies in $\mathbf{1}^{\perp_\mu}$.

Using $V = \mathbf{1}m + V_c$ and $A\mathbf{1} = \mathbf{1}$, we obtain

$$AV = A(\mathbf{1}m + V_c) = \mathbf{1}m + AV_c. \tag{49}$$

This proves the additive decomposition.

We now prove the orthogonality of the two terms in $L^2(\pi)$. Since $\pi^\top A = \mu^\top$ and $\mu^\top V_c = 0$,

$$\langle \mathbf{1}m, AV_c\rangle_{\pi,F} = \operatorname{tr}\big((\mathbf{1}m)^\top \Omega_\pi AV_c\big) = \operatorname{tr}\big(m^\top \mathbf{1}^\top \Omega_\pi AV_c\big) = \operatorname{tr}\big(m^\top \pi^\top AV_c\big) = \operatorname{tr}\big(m^\top \mu^\top V_c\big) = 0. \tag{50}$$

Finally, write $V_c = [v_c^{(1)}, \ldots, v_c^{(d)}]$, where each $v_c^{(r)} \in \mathbf{1}^{\perp_\mu}$. Applying Corollary 1 to each feature column gives

$$\|Av_c^{(r)}\|^2_\pi \le \lambda_2(D_{\mu,\pi})\|v_c^{(r)}\|^2_\mu. \tag{51}$$

Summing over feature coordinates,

$$\|AV_c\|^2_{\pi,F} = \sum_{r=1}^{d} \|Av_c^{(r)}\|^2_\pi \le \lambda_2(D_{\mu,\pi}) \sum_{r=1}^{d} \|v_c^{(r)}\|^2_\mu = \lambda_2(D_{\mu,\pi})\|V_c\|^2_{\mu,F}. \tag{52}$$

Taking square roots gives

$$\|AV_c\|_{\pi,F} \le \sqrt{\lambda_2(D_{\mu,\pi})}\,\|V_c\|_{\mu,F}. \tag{53}$$

A.7 Proof of Corollary 2

Let

$$Z = \Omega_\pi^{1/2} AV, \qquad Z_c = \Omega_\pi^{1/2} AV_c. \tag{54}$$

By Proposition 3,

$$AV = \mathbf{1}m + AV_c. \tag{55}$$

Multiplying by $\Omega_\pi^{1/2}$ gives

$$Z = \Omega_\pi^{1/2}\mathbf{1}m + Z_c. \tag{56}$$

Define

$$Z_m := \Omega_\pi^{1/2}\mathbf{1}m. \tag{57}$$

Since π is a probability measure,

$$Z_m^\top Z_m = m^\top \mathbf{1}^\top \Omega_\pi \mathbf{1}m = m^\top m. \tag{58}$$

Moreover, the cross term vanishes as a matrix:

$$Z_m^\top Z_c = m^\top \mathbf{1}^\top \Omega_\pi AV_c = m^\top \pi^\top AV_c = m^\top \mu^\top V_c = 0. \tag{59}$$

Therefore

$$Z^\top Z = Z_m^\top Z_m + Z_c^\top Z_c = m^\top m + Z_c^\top Z_c, \tag{60}$$

and

$$\|Z\|^2_F = \|m\|^2_2 + \|Z_c\|^2_F. \tag{61}$$

Assume first that both $m \neq 0$ and $Z_c \neq 0$. Define the normalized spectral densities

$$\rho_Z = \frac{Z^\top Z}{\operatorname{tr}(Z^\top Z)}, \qquad \rho_m = \frac{m^\top m}{\|m\|_2^2}, \qquad \rho_c = \frac{Z_c^\top Z_c}{\|Z_c\|_F^2}. \tag{62}$$

Let

$$\alpha = \frac{\|Z_c\|_F^2}{\|m\|_2^2 + \|Z_c\|_F^2}. \tag{63}$$

Then

$$\rho_Z = (1-\alpha)\rho_m + \alpha\rho_c. \tag{64}$$

The matrix ρ_m has rank one, so its entropy is zero:

$$H(\rho_m) = 0. \tag{65}$$

Using concavity of entropy,

$$H(\rho_Z) \geq (1-\alpha)H(\rho_m) + \alpha H(\rho_c) = \alpha H(\rho_c). \tag{66}$$

Since $H(\rho_Z) = H(Z)$ and $H(\rho_c) = H(Z_c)$, this gives

$$H(Z) \geq \alpha H(Z_c). \tag{67}$$

For the upper bound, we use the standard entropy bound for a two-component mixture:

$$H\big((1-\alpha)\rho_m + \alpha\rho_c\big) \leq h_2(\alpha) + (1-\alpha)H(\rho_m) + \alpha H(\rho_c), \tag{68}$$

where

$$h_2(\alpha) = -\alpha\log\alpha - (1-\alpha)\log(1-\alpha). \tag{69}$$

Since $H(\rho_m) = 0$, this yields

$$H(Z) \leq h_2(\alpha) + \alpha H(Z_c). \tag{70}$$

The cases $m = 0$ or $Z_c = 0$ follow by continuity, with the convention $\alpha H(Z_c) = 0$ when $Z_c = 0$.

Finally,

$$\|Z_c\|_F = \|\Omega_\pi^{1/2} A V_c\|_F = \|A V_c\|_{\pi,F}. \tag{71}$$

The routed contraction bound from Proposition 3 therefore gives

$$\|Z_c\|_F \leq \sqrt{\lambda_2(D_{\mu,\pi})}\,\|V_c\|_{\mu,F}. \tag{72}$$

A.8 Centered participation ratio

To isolate the nontrivial routed directions, we may also define the centered participation ratio

$$\mathrm{PR}_{\mathrm{cent}}(D_{\mu,\pi}) = \frac{1}{n-1}\frac{\left(\sum_{k=2}^n \lambda_k\right)^2}{\sum_{k=2}^n \lambda_k^2}, \tag{73}$$

whenever $\sum_{k=2}^n \lambda_k^2 > 0$. If all centered eigenvalues vanish, we set $\mathrm{PR}_{\mathrm{cent}}(D_{\mu,\pi}) = 0$.

B Additional Results

B.1 Correlation on other models.

We report the version of Figure 3 but using three different models, GPT-2, Pythia-410m [6] and Qwen2-7B [44]. The correlation plots are provided in Figure 5. We note that the relation between the metrics is stable across models, being based on structural properties common to all attention maps.

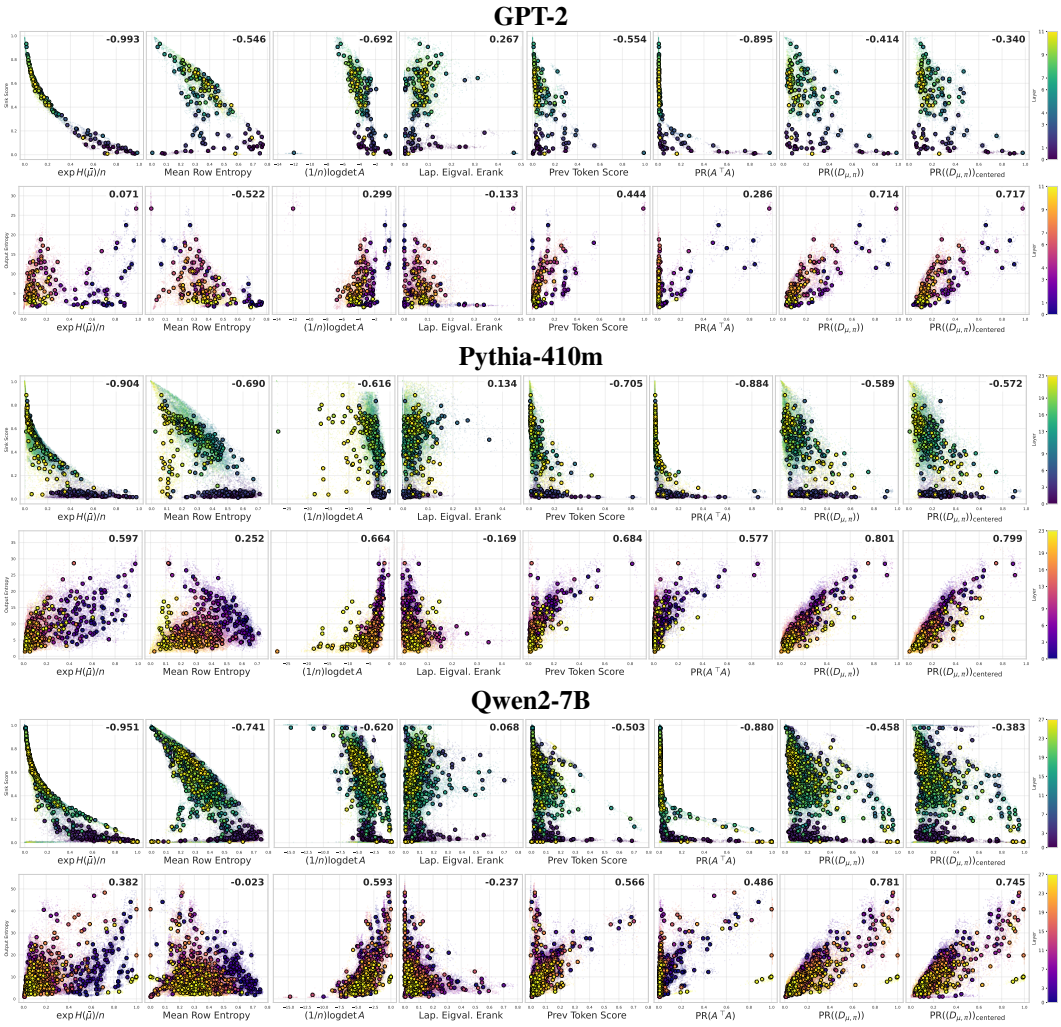


Figure 5: Scatterplot between attention diagnostics, sink score, and output dimensionality on GPT-2, Pythia-410m and Qwen2-7B

Table 3: Supervised hallucination-detection performance (ROC-AUC). We compare attention-matrix baselines, Laplacian eigenvalues, and features derived from the weighted token-difference operator. We follow the experimental protocol of [7] using Llama-3.1-8B [1], DeepSeek-4 as a judge.

Dataset	AttnEigvals	AttnLogDet	LapEigvals	TokenDiffEigvals
CoQA	0.7995	0.7779	**0.8519**	<u>0.8241</u>
GSM8K [10]	0.8598	0.8586	<u>0.8660</u>	**0.8663**
TriviaQA	0.8385	0.8346	**0.8923**	<u>0.8731</u>
TruthfulQA	0.7997	<u>0.8362</u>	0.8095	**0.8445**

B.2 Supervised hallucination detection with spectral probes

Beyond unsupervised SAUQ, we probe the token-difference spectrum in the supervised hallucination-detection protocol of LapEigval [7], which trains a classifier on eigenvalue features of the attention-induced graph Laplacian. We replace those features with the leading $k = 16$ eigenvalues of $D_{\mu,\pi}$ for each selected head. Truncation is necessary because, unlike the Laplacian, whose diagonal already supplies cheap summaries, $D_{\mu,\pi}$ is dense and weighted, so its full spectrum is comparatively expensive to estimate.

Estimator	Category	QA					Summarization		
		TriviaQA	TruthfulQA	MMLU	CoQA	Mean	SAMSum	XSum	Mean
AttentionScore [36]	Single pass	0.0429	0.0466	0.1114	0.0022	0.0508	0.2924	-0.0540	0.1192
AttentionScore (gen. only) [36]	Single pass	0.1529	-0.1053	0.6741	0.0892	0.2027	-0.0726	-0.5038	-0.2882
MaximumSequenceProbability	Single pass	0.2328	0.0436	0.6644	0.3016	0.3106	0.1220	-0.0610	0.0305
MeanTokenEntropy	Single pass	0.6173	0.2474	0.2491	0.2533	0.3418	<u>0.4265</u>	0.3363	**0.3814**
Perplexity	Single pass	-0.0126	0.2093	0.4042	0.2620	0.2157	**0.4303**	0.3309	<u>0.3806</u>
CCP [16]	Single pass	0.0636	-0.0244	0.4957	0.2334	0.1921	0.0975	0.1419	0.1197
NLI-DegMat [26]	Sample-based	**0.6393**	0.3117	0.3758	**0.4073**	<u>0.4335</u>	0.2956	0.3367	0.3162
NLI-Eccentricity [26]	Sample-based	0.6149	0.2472	0.3318	0.3476	0.3854	0.2883	0.2600	0.2742
NLI-EigValLaplacian [26]	Sample-based	0.6201	<u>0.3147</u>	0.3213	0.3855	0.4104	0.2934	<u>0.3535</u>	0.3235
EigenScore [9]	Sample-based	0.4990	-0.0506	0.3184	0.3282	0.2738	0.1783	0.0123	0.0953
LUQ [46]	Sample-based	<u>0.6334</u>	**0.3777**	0.3576	<u>0.3900</u>	**0.4397**	0.2697	**0.4477**	0.3587
LexicalSimilarity_rougeL [17]	Sample-based	0.6001	0.1317	0.3628	0.3697	0.3661	0.3237	0.2437	0.2837
SAR [14]	Sample-based	0.6116	0.2120	0.4401	0.3446	0.4021	0.2869	0.3027	0.2948
SemanticDensity [31]	Sample-based	0.4903	0.1312	0.1323	0.0731	0.2067	0.1083	0.1898	0.1491
SemanticEntropy [23]	Sample-based	0.3098	0.1179	0.3018	0.3066	0.2590	0.2640	0.0225	0.1433
RAUQ [40]	Single pass	0.5226	0.1970	0.5939	0.2639	0.3944	0.3898	0.2750	0.3324
SAUQ (gen. only)	Single pass	0.5906	0.1290	0.5182	0.2953	0.3833	0.3824	0.3484	0.3654
SAUQ (full PR)	Single pass	0.3780	0.1804	**0.7204**	0.2107	0.3724	0.3988	0.3165	0.3577
SAUQ (Ours)	Single pass	0.3996	0.1817	<u>0.7176</u>	0.2115	0.3776	0.4005	0.3211	0.3608

Table 4: Normalized PRR (higher is better) of unsupervised uncertainty estimators on Qwen2.5-7B. Sample-based methods use 10 samples. The best result in each column is shown in bold, and the second best is underlined.

Table 3 shows that TDO eigenvalues are competitive with Laplacian features. In the low-k regime, the leading eigenvalues of $D_{\mu,\pi}$ often match or exceed the Laplacian features, suggesting that the dominant routing modes carry hallucination-relevant information. We do not claim uniform dominance: with richer or cheaper Laplacian features, the baseline can match or surpass the probe. Together with the unsupervised SAUQ results, this suggests that the $D_{\mu,\pi}$ spectrum carries uncertainty information not reducible to scalar attention statistics.

B.3 Uncertainty quantification on Qwen2.5-7B

Table 4 replicates the evaluation of Section 5 on Qwen2.5-7B [32], using the same protocol, the same $\alpha = 0.2$ and the middle third of layers.

The results confirm the regime dependence of Section 5, but they also show that the gains are model dependent. SAUQ improves over RAUQ on summarization (0.3608 vs. 0.3324) and, together with its full-PR variant, obtains the best MMLU scores (0.7176 and 0.7204), but it is below RAUQ on the QA mean (0.3776 vs. 0.3944). The centered and full PR signals differ by at most 0.022 on every dataset. Unlike on Llama-3.1-8B, probability-based estimators (MeanTokenEntropy, Perplexity) are the strongest on summarization, and the generated-only variant outperforms the full-context one (0.3654 vs. 0.3608). We interpret the smaller advantage of the full-context signal as consistent with weaker sink behavior in this model, since the intrinsic measure primarily corrects sink-induced bias; we list the dependence on architecture as a limitation in Section 6.

B.4 Confidence intervals

We quantify the variability of PRR with a nonparametric bootstrap over evaluation instances: for each dataset, we resample instances with replacement 5000 times, recompute PRR for every estimator on the same resample, and report the standard errors. Differences between estimators are computed on paired resamples. Generation is greedy, so there is no decoding randomness in single-pass estimators; sample-based baselines use the 10 samples provided by LM-Polygraph. Table 5 reports the intervals for the attention-based estimators and the strongest baselines.

B.5 Calibration

PRR measures the quality of the ranking induced by an uncertainty score, and SAUQ is designed as a ranking score so it is not a probability. Table 6 reports the expected calibration error of single-pass estimators over the evaluation set, with 20 equal-width bins.

Estimator	Category	QA				Summarization	
		TriviaQA	TruthfulQA	MMLU	CoQA	SAMSum	XSum
MaximumSequenceProbability	Single pass	$0.25 \pm .02$	$0.32 \pm .06$	**0.55** $\pm .02$	$0.29 \pm .02$	$-0.14 \pm .07$	$-0.39 \pm .11$
MeanTokenEntropy	Single pass	$0.57 \pm .02$	**0.42** $\pm .05$	$0.24 \pm .03$	$0.26 \pm .02$	$0.50 \pm .05$	**0.39** $\pm .08$
Perplexity	Single pass	$0.11 \pm .02$	$0.29 \pm .05$	$0.52 \pm .02$	$0.27 \pm .02$	$0.51 \pm .04$	**0.41** $\pm .08$
CCP [16]	Single pass	$0.10 \pm .02$	$0.16 \pm .06$	**0.56** $\pm .02$	$0.26 \pm .02$	$-0.15 \pm .07$	$-0.30 \pm .11$
NLI-DegMat [26]	Sample-based	**0.58** $\pm .02$	$0.21 \pm .06$	$0.44 \pm .02$	**0.40** $\pm .02$	$0.39 \pm .05$	**0.32** $\pm .09$
NLI-Eccentricity [26]	Sample-based	$0.56 \pm .02$	$0.20 \pm .06$	$0.45 \pm .02$	$0.38 \pm .02$	$0.34 \pm .05$	$0.04 \pm .10$
NLI-EigValLaplacian [26]	Sample-based	**0.60** $\pm .02$	$0.20 \pm .06$	$0.47 \pm .02$	**0.41** $\pm .02$	$0.39 \pm .05$	**0.38** $\pm .09$
EigenScore [9]	Sample-based	$0.36 \pm .02$	$0.05 \pm .06$	$0.41 \pm .03$	$0.35 \pm .02$	$0.07 \pm .06$	$-0.21 \pm .10$
LUQ [46]	Sample-based	**0.59** $\pm .02$	$0.18 \pm .06$	$0.44 \pm .02$	$0.36 \pm .02$	$0.29 \pm .05$	**0.36** $\pm .08$
LexicalSimilarity (ROUGE-L) [17]	Sample-based	$0.48 \pm .02$	$0.16 \pm .06$	$0.41 \pm .03$	$0.36 \pm .02$	$0.25 \pm .06$	$0.20 \pm .09$
SAR [14]	Sample-based	$0.57 \pm .02$	$0.18 \pm .06$	$0.41 \pm .03$	$0.33 \pm .02$	$0.29 \pm .06$	**0.31** $\pm .09$
SemanticDensity [31]	Sample-based	$0.56 \pm .02$	$0.08 \pm .06$	$0.16 \pm .03$	$0.11 \pm .02$	$0.08 \pm .06$	$0.11 \pm .10$
SemanticEntropy [23]	Sample-based	$0.20 \pm .03$	$0.21 \pm .06$	$0.47 \pm .02$	$0.37 \pm .02$	$0.25 \pm .06$	$-0.19 \pm .11$
AttentionScore [36]	Single pass	$0.08 \pm .02$	$-0.05 \pm .06$	$0.21 \pm .03$	$0.09 \pm .02$	$-0.37 \pm .07$	$-1.00 \pm .11$
RAUQ [40]	Single pass	$0.26 \pm .02$	$0.34 \pm .05$	**0.56** $\pm .02$	$0.31 \pm .02$	$0.53 \pm .05$	**0.30** $\pm .08$
LapEigvalRAUQ [7]	Single pass	$0.12 \pm .02$	$0.33 \pm .06$	**0.56** $\pm .02$	$0.31 \pm .02$	$0.53 \pm .04$	$0.29 \pm .08$
LogDetRAUQ [36]	Single pass	$0.11 \pm .02$	$0.33 \pm .05$	**0.56** $\pm .02$	$0.31 \pm .02$	$0.53 \pm .04$	$0.30 \pm .08$
SAUQ (gen. only)	Single pass	$0.50 \pm .02$	$0.31 \pm .05$	**0.54** $\pm .02$	$0.31 \pm .02$	$0.49 \pm .05$	$0.26 \pm .09$
SAUQ (full PR)	Single pass	$0.29 \pm .02$	**0.38** $\pm .05$	**0.56** $\pm .02$	$0.31 \pm .02$	**0.59** $\pm .04$	**0.42** $\pm .07$
SAUQ (Ours)	Single pass	$0.27 \pm .02$	**0.39** $\pm .05$	**0.56** $\pm .02$	$0.30 \pm .02$	**0.60** $\pm .04$	**0.42** $\pm .07$

Table 5: Normalized PRR (higher is better) of unsupervised uncertainty estimators on Llama-3.1-8B, mean $\pm$ standard error (stratified bootstrap over evaluation instances, 5000 resamples; one greedy decode per instance, 10 samples for sample-based methods). Bold: not significantly worse than the best result in the column.

Estimator	QA (mean ECE)	Summ. (mean ECE)
Perplexity	0.237	0.558
AttentionScore	0.266	0.551
MaxSeqProbability	0.265	0.685
MeanTokenEntropy	**0.162**	0.505
RAUQ	0.187	0.522
SAUQ (full PR)	0.194	**0.490**
SAUQ (Ours)	0.190	0.498

Table 6: Expected calibration error (lower is better) of single-pass estimators on Llama-3.1-8B, averaged over the QA and summarization datasets.

B.6 Head selection and routing signal

SAUQ enters the estimator at two places: the rule that selects one head per layer, and the signal computed on that head inside the recurrence. To separate the two, we vary them independently on TruthfulQA and SAMSum (Table 7). Rows are selection rules, columns are signals; (Gen) denotes the generated-token block, (Full) the full prefix, and (Eucl) and (Intrinsic) the counting and intrinsic geometries. Our configuration is generated-Euclidean selection with the centered full-intrinsic signal.

Selection $\downarrow$ / Signal $\rightarrow$	gen-Eucl	full-Eucl	full-intr.	full-intr.-cent.
TruthfulQA				
gen-Euclidean	0.3056	0.3494	0.3767	**<u>0.3876</u>**
full-Euclidean	0.3203	0.3507	0.3761	0.3759
full-intrinsic	0.3078	0.3508	0.3482	0.3429
full-intr.-cent.	0.3093	0.3508	0.3482	0.3429
SAMSum				
gen-Euclidean	0.4948	0.5384	0.5924	**<u>0.6028</u>**
full-Euclidean	0.4398	0.5375	0.5358	0.5371
full-intrinsic	0.4916	0.5371	0.5144	0.5339
full-intr.-cent.	0.5021	0.5373	0.5054	0.5335

Table 7: PRR for every combination of head-selection rule (rows) and routing signal (columns). Underlined: the configuration reported in Table 1; bold: best per dataset.

Three observations follow. First, the benefit of the intrinsic geometry on the signal axis depends on the selection rule: under our generated-Euclidean selection, replacing the full-Euclidean signal with the full-intrinsic one improves PRR on both datasets ($0.3494 \to 0.3767$ and $0.5384 \to 0.5924$), but across all eight selection rows it does so in only three. Second, the centered PR improves over the full PR under our selection rule on both datasets, and in five of the eight rows overall (all four on SAMSum, one on TruthfulQA); the gain is therefore not uniform. Third, the intrinsic geometry does not help on the selection axis, which is why SAUQ keeps the Euclidean proxy for selection.

Cost of intrinsic selection. Selection screens all H heads in every monitored layer, whereas the signal is computed only for the selected head. Intrinsic selection would require the TDO of the full $(m+T) \times (m+T)$ matrix for every head, i.e. H times the cost of the signal, while the Euclidean proxy only uses the $T \times T$ generated block.

B.7 Length-controlled analysis

Summarization and QA differ in task, input length, and output length, so the summarization gains of Section 5 could reflect any of these. We stratify the difference $\Delta = \mathrm{PRR}_{\mathrm{SAUQ}} - \mathrm{PRR}_{\mathrm{RAUQ}}$ by input and output length on Llama-3.1-8B, using a stratified bootstrap with 5,000 resamples. Table 8 reports the pooled difference and the difference after controlling for output length, and for input and output length jointly.

Dataset	Task	Δ pooled	95% CI	ctrl. output	ctrl. input$\times$output
XSum	Summ.	$+0.116$	$[+0.040, +0.199]$	$+0.076$	$+0.056$
SAMSum	Summ.	$+0.076$	$[+0.018, +0.135]$	$+0.013$	$+0.026$
TruthfulQA	QA	$+0.051$	$[-0.001, +0.099]$	$+0.041$	$+0.034$
TriviaQA	QA	$+0.007$	$[-0.024, +0.039]$	$+0.026$	$+0.038$
CoQA	QA	-0.006	$[-0.016, +0.003]$	-0.013	-0.010
MMLU	QA	-0.002	$[-0.006, +0.002]$	-0.015	-0.021

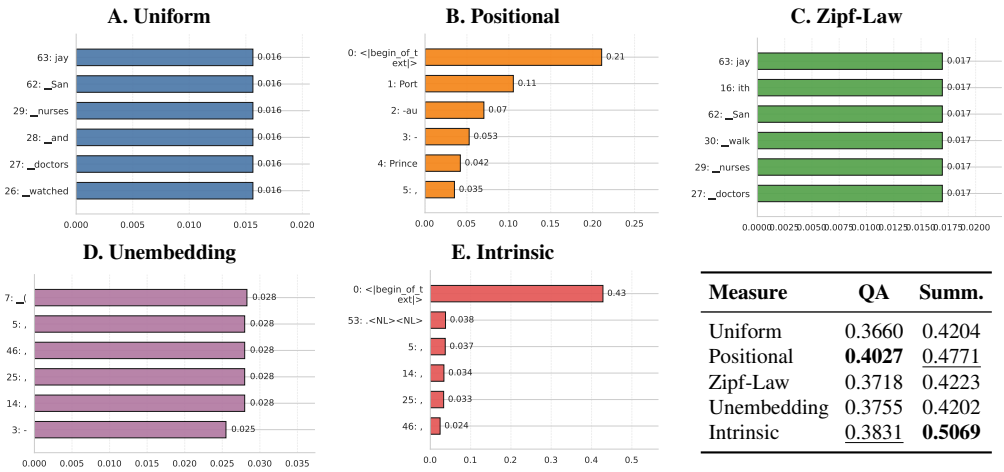
Table 8: Difference in PRR between SAUQ and RAUQ on Llama-3.1-8B, pooled and after stratifying by output length and by input $\times$ output length.

Roughly half of the summarization gain is explained by generation length: the summarization mean of Δ drops from 0.096 to 0.044 under the output-length control, while the input-length control leaves it nearly unchanged (0.089). The residual is carried by XSum, where it remains positive under both controls ($+0.076$ and $+0.056$, although the controlled intervals include zero) and grows with context: across input-length quartiles it is $+0.006$, $+0.053$, $+0.142$ and $+0.194$, the last significantly positive. On SAMSum, whose inputs span a narrower range of lengths, the advantage does not survive the output-length control ($+0.076 \to +0.013$).

B.8 Choice of measure

Each measure ν below replaces the input measure μ in the TDO, giving $D_{\nu,\pi} = \Omega_\nu^{-1} A^\top \Omega_\pi A$, and is normalized to sum to one over the n tokens of the prefix. In all cases $\pi = \frac{1}{n}\mathbf{1}$. In Figure 6, we show an example on how the mass are distributed.

- **Uniform:** $\nu_j = 1/n$.
- **Positional:** $\nu_l = l^{-\gamma}$ with $\gamma = 1$. Early positions, including the sink, receive large mass; since keys enter $D_{\nu,\pi}$ through Ω_ν^{-1}, their columns are down-weighted.
- **Zipf-Law (in-context frequency):** $\nu_l = \big|\{k \le N : x_k = x_l\}\big|$, the number of occurrences of token x_l in the sequence. Tokens that occur once receive the same mass, so on short inputs this measure is close to uniform.
- **Unembedding:** $\nu_l = w_{x_l}$ with $w_t = \exp\big(\sigma p_t - \max_{t'} \sigma p_{t'}\big)$, where $p_t = (W_U[t] - \bar{w})^\top v_1$, $\bar{w}$ is the mean row of W_U, v_1 the top right singular vector of the centered W_U, and $\sigma \in \{\pm 1\}$ is chosen so that σp_t is negatively correlated with the token id.
- **Intrinsic:** $\nu = \mu = A^\top \pi$, the average attention row of the head.



Measure	QA	Summ.
Uniform	0.3660	0.4204
Positional	**0.4027**	<u>0.4771</u>
Zipf-Law	0.3718	0.4223
Unembedding	0.3755	0.4202
Intrinsic	<u>0.3831</u>	**0.5069**

Figure 6: Top tokens receiving the largest probability mass under each measure, and mean PRR for QA and summarization of SAUQ with these masses. Input (first 100 chars): *Port-au-Prince, Haiti (CNN) – Earthquake victims, writhing in pain and grasping at life, watched doctors and nurses....*

B.9 Resource details.

All experiments were run on a single NVIDIA L40S GPU with 48 GB of memory, sufficient to hold 8-billion-parameter models. Running the full SAUQ evaluation on the unsupervised benchmark takes approximately 8 hours per model. The routing signal requires only $\mathrm{tr}(S)$ and $\mathrm{tr}(S^2) = \|S\|_F^2$ of the symmetric matrix $S_{\mu,\pi}$, so no eigendecomposition is needed; forming $S_{\mu,\pi}$ costs $\mathcal{O}(n_t^3)$ per selected head and step with dense products.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's contributions: an operator-theoretic framework for attention (Sec. 3) that characterizes the relation between sinking and head output representation (Sec 4), and a method for uncertainty estimation (Sec 5).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper presents a paragraph in Sec 6 stating the main limitations of the proposed approach.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Sec 3 states all the necessary assumptions to prove the results. Proof sketches are included in the main text, with full proofs of the propositions and corollaries provided in Supplementary A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Sec 5, the paper specifies SAUQ recurrence, layer/head selection rule, generated-token variant, model, benchmark, datasets, metric, baselines, and the use of 10 samples for sample-based methods. The experiments use public benchmarks and follow the LM Polygraph evaluation protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is released under the MIT license at `https://github.com/gviga/spectral-geometry-attention`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the main evaluation setup, including Llama-3.1-8B, the LM Polygraph benchmark, PRR as the evaluation metric, the QA and summarization datasets, the compared baselines, the spectral head-selection procedure, and $\alpha = 0.2$ for SAUQ. No optimizer is required for the main unsupervised SAUQ experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The Supplementary material reports standard errors for all PRR scores on Llama-3.1-8B from a nonparametric bootstrap over evaluation instances (5000 resamples, paired across estimators), and 95% bootstrap confidence intervals for the length-controlled SAUQ–RAUQ differences.

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper discloses the hardware necessary to replicate experiments, especially in terms of Gb of VRAM, in the Supplementary Materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work analyzes existing LLMs and public benchmarks for attention interpretability and uncertainty estimation, and does not introduce human-subject experiments, private data collection, or high-risk model release. The authors have reviewed the NeurIPS Code of Ethics and believe the work conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper motivates positive uses for uncertainty estimation and interpretability methods in a dedicated paragraph in Sec 6.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new high-risk pretrained model, image generator, scraped dataset, or other asset requiring controlled release safeguards. The work proposes analysis and uncertainty-estimation methods evaluated on existing models and benchmarks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the existing model, benchmarks, datasets, and baseline methods used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only new asset is the code package, released under the MIT license with a README documenting installation and usage. No new dataset or pretrained model is released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not conduct new crowdsourcing experiments or new human-subject studies. It evaluates on existing public benchmarks and datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve new crowdsourcing or human-subject research conducted by the authors. Therefore, IRB or equivalent approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are central to the research: the paper analyzes transformer attention and evaluates SAUQ using Llama-3.1-8B and Qwen2.5-7B on uncertainty-estimation benchmarks. The model usage is described in Sections 4 and 5 as part of the core methodology and evaluation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.