A HARMONIC BASIS FOR AUDITING LEARNED SEQUENTIAL BIASES IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we show that the next-token statistics of any autoregressive language model, when observed at a chosen context order k, induce a weighted random walk on the de Bruijn graph G_k. We develop a harmonic analysis of this induced walk using the recently derived closed-form eigenvectors of the de Bruijn graph Laplacian. Each eigenvector captures an independent pattern of token co-occurrence at a definite oscillatory scale, and projecting the model's stationary distribution onto this basis yields temperature-independent coefficients – the model's *harmonic spectrum* – that depend only on the model's centred logits and account exactly for every deviation from uniform generation. We validate the framework with several experiments. First, the harmonic coordinates reveal distinct sequential biases in the task of binary random sequence generation with LLMs. Second, the harmonic spectrum decomposes the exploitable structure of LLM strategies in Rock-Paper-Scissors game into interpretable components, and correlates strongly with win rate across 19 models. Finally, we apply our framework to open-ended generation, projecting part-of-speech tag sequences onto the harmonic basis.

1 INTRODUCTION

Autoregressive language models generate text through a deceptively simple recurrence: at each step, sample the next token from a learned distribution over the current context window, discard the oldest token, and append the new one. This shift-register update – so named for its use in hardware sequence generators – is the defining structural operation of de Bruijn graphs (de Bruijn, 1946; Good, 1946). The de Bruijn graph G_k of order k over an alphabet $\mathcal{A}$ has $(k-1)$-grams as vertices and k-grams as directed edges; the shift-register update is exactly an edge traversal. We observe that a language model with context length $k-1$ therefore defines a weighted random walk on G_k, where the weight of edge (u, v) is the model's conditional probability of producing token v_{k-1} given context u.

This identification, which has not been made explicit in the literature, opens an unexpected door. A recent result of Philippakis et al. (Philippakis et al., 2024) provides, for the first time, closed-form eigenvectors for the de Bruijn graph Laplacian, expressed in terms of a discrete Fourier transform over $\mathcal{A}$ and a simple padding operation. These eigenvectors form a natural harmonic basis for functions on contexts. The present work asks: *what does this spectral structure tell us about a language model?*

Our contributions. Our answer is that the stationary distribution of an autoregressive model admits a Fourier decomposition on G_k, in precise analogy with decomposing a function into sinusoidal modes. Performing the same Fourier transform used by Philippakis et al. over the vocabulary diagonalises the Laplacian into a block-diagonal form, with each block corresponding to an independent Fourier mode – a specific pattern of token co-occurrence at a definite oscillatory scale. The projection of the model's stationary distribution onto each mode yields a temperature-independent coefficient γ_i, and together these coefficients form the model's *harmonic spectrum*: a compact summary of model's sequential biases. The contributions of this work are the following.

1. We identify autoregressive inference as a weighted random walk on the de Bruijn graph G_k, connecting LLMs to the closed-form Laplacian spectral theory of de Bruijn graphs (§2–3).
2. We prove the stationary distribution admits a temperature-invariant Fourier decomposition: $\langle \pi^{(T)} - \pi^{(\infty)}, x_i \rangle = \gamma_i / T + O(1/T^2)$, where γ_i is fixed by the model's centred logits through a Poisson equation and vanishes iff those logits are balanced over incoming contexts (§4).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

3. Empirical validations. (i) In binary sequence generation (§5.1), the harmonic coordinates cleanly separate the three structured regimes and transfer from abstract $\{0, 1\}$ to gender tokens. (ii) In Rock-Paper-Scissors (§5.2), the harmonic spectrum decomposes exploitable play into interpretable components and predicts win rate across 19 models at $r = -0.89$.

2 Background: De Bruijn Graphs and Language Model Inference

2.1 De Bruijn Graphs

Definition 2.1 (Alphabet and k-mers). *Let $\mathcal{A}$ be a finite alphabet of cardinality $q = |\mathcal{A}|$. For integer $k \geq 1$, a k-*mer *is any string $s = s_0 s_1 \cdots s_{k-1} \in \mathcal{A}^k$.*

Definition 2.2 (De Bruijn graph). *The* de Bruijn graph of order k, *denoted $G_k = (\mathcal{A}^{k-1}, \mathcal{A}^k)$, has vertices given by all $(k-1)$-mers $u \in \mathcal{A}^{k-1}$, and directed edges given by all k-mers $e = s_0 s_1 \cdots s_{k-1} \in \mathcal{A}^k$, where e points from $u = s_0 \cdots s_{k-2}$ to $v = s_1 \cdots s_{k-1}$. Thus $|V(G_k)| = q^{k-1}$ and $|E(G_k)| = q^k$.*

The defining structural property is the *overlap condition*: the last $k-2$ letters of the source vertex equal the first $k-2$ letters of the target. This is precisely the shift-register recurrence.

Convention. Throughout this paper we set k to equal the context length plus one: a model that conditions on L previous tokens operates on G_{L+1}, with vertices being L-grams and edges being $(L+1)$-grams. We write $k-1$ for the context length wherever it is needed explicitly. (Thus the de Bruijn order k is always one greater than the language model's context window.)

2.2 Language Model Inference as a Shift-Register Walk

Consider an autoregressive language model $\mathcal{M}$ with context length $k-1$ and vocabulary $\mathcal{A}$. For context $u = s_0 \cdots s_{k-2} \in \mathcal{A}^{k-1}$, the model produces a distribution $P_{\mathcal{M}}(\cdot \mid u)$ over next tokens.

Notation. For a string $s = s_0 s_1 \cdots s_{k-1}$, write $s_{1:} = s_1 \cdots s_{k-1}$ (drop s_0) and $s_{:k-2} = s_0 \cdots s_{k-2}$ (drop s_{k-1}). For example, for $s = \texttt{0110}$, $s_{1:} = \texttt{110}$ and $s_{:k-2} = \texttt{011}$.

The generation process

$$s_t \sim P_T(\cdot \mid s_{t-k+1} \cdots s_{t-1}), \qquad (s_{t-k+1} \cdots s_{t-1}) \;\to\; (s_{t-k+2} \cdots s_t) \tag{1}$$

is a random walk on G_k: the walker at vertex u moves along edge $e = u \cdot s_t$ to vertex $v = s_{t-k+2} \cdots s_t$ with probability $P_{\mathcal{M}}(s_t \mid u)$. For the first $k-1$ steps, the model uses a fixed prompt or special start-of-sequence padding; the stationary analysis is independent of this initialisation.

Definition 2.3 (LLM random walk on G_k). *The transition probabilities of the LLM-induced random walk on $V(G_k) = \mathcal{A}^{k-1}$ are*

$$T_{u \to v} = P_{\mathcal{M}}(v_{k-2} \mid u) \;\; \textit{if} \;\; u_{1:} = v_{:k-2}, \qquad T_{u \to v} = 0 \;\; \textit{otherwise,} \tag{2}$$

Temperature scaling. The softmax temperature T rescales the model's logits $\ell(s \mid u)$:

$$P_T(s \mid u) = \frac{\exp(\ell(s \mid u)/T)}{\sum_{s'} \exp(\ell(s' \mid u)/T)}. \tag{3}$$

Because the de Bruijn graph G_k is strongly connected (de Bruijn, 1946) and the softmax assigns positive probability to every outgoing edge, the induced random walk is irreducible and aperiodic, so a unique stationary distribution $\pi^{(T)}$ with a transition matrix $T^{(T)}$ exist (Zekri et al., 2024). In the limit $T \to \infty$, every transition probability converges to $1/q$; we write $T^{(\infty)}$ for this limiting transition matrix, defined entry-wise by

$$T^{(\infty)}_{u \to v} = \begin{cases} 1/q & \text{if } u_{1:} = v_{:k-2}, \\ 0 & \text{otherwise,} \end{cases}$$

and $\pi^{(\infty)} = q^{-(k-1)}\mathbf{1}$ for its (uniform) stationary distribution.

3 THE SPECTRAL TRANSFORM: HARMONIC MODES OF LANGUAGE

We now construct the harmonic basis for functions on contexts. The key idea is to perform a Fourier transform over the vocabulary $\mathcal{A}$, which simultaneously block-diagonalises the graph Laplacian and assigns every basis vector an explicit linguistic interpretation. The mathematical foundations are due to Philippakis et al. (Philippakis et al., 2024); our contribution here is to foreground the harmonic-analysis reading throughout.

3.1 THE TENSOR STRUCTURE OF CONTEXT SPACE

A context is a sequence of $k-1$ tokens. The space of all functions on contexts is naturally the $(k-1)$-fold tensor product of functions on single tokens.

Definition 3.1 (Token space and context space)**.** *Let $V_{\mathcal{A}} = \mathbb{C}^q$ (the q-dimensional vector space over the complex numbers) with standard basis indexed by $\mathcal{A}$. The space of all functions on contexts is $V_{\mathcal{A}}^{\otimes(k-1)}$, and the space of all functions on k-grams is $V_{\mathcal{A}}^{\otimes k}$.*

Definition 3.2 (Deletion operators (Philippakis et al., 2024))**.** *For $k \geq 2$, the* left and right deletion operators $\theta_L, \theta_R : V_{\mathcal{A}}^{\otimes k} \to V_{\mathcal{A}}^{\otimes(k-1)}$ *act on pure tensors by*

$$\theta_L(v_0 \otimes \cdots \otimes v_{k-1}) = q^{-1/2}\, v_1 \otimes \cdots \otimes v_{k-1}, \qquad \theta_R(v_0 \otimes \cdots \otimes v_{k-1}) = q^{-1/2}\, v_0 \otimes \cdots \otimes v_{k-2}. \tag{4}$$

Their adjoints (up to the constant factor $q^{-1/2}$) $\theta_L^*, \theta_R^* : V_{\mathcal{A}}^{\otimes(k-1)} \to V_{\mathcal{A}}^{\otimes k}$ prepend and append the de Bruijn vector $\bullet = q^{-1/2} \sum_{a \in \mathcal{A}} a$*:*

$$\theta_L^*(v) = \bullet \otimes v, \qquad \theta_R^*(v) = v \otimes \bullet. \tag{5}$$

Linguistic interpretation of the operators. θ_L strips the first (oldest) token from a k-gram; θ_R strips the last (newest). Their adjoints pad a context with $\bullet$, the *average token*: the uniform superposition over all letters. In operator language, one step of autoregressive generation maps a context $u \in V_{\mathcal{A}}^{\otimes(k-1)}$ to a new context $v \in V_{\mathcal{A}}^{\otimes(k-1)}$ by (i) forming the k-gram $u \otimes e_s$ (appending the sampled token s) and then (ii) applying θ_L to discard the oldest token: $v = \theta_L(u \otimes e_s)$. The composition $\theta_L(\cdot \otimes e_s) : V_{\mathcal{A}}^{\otimes(k-1)} \to V_{\mathcal{A}}^{\otimes(k-1)}$ is a single edge traversal on G_k. The de Bruijn vector $\bullet$ is the *zero-mode*: in $V_{\mathcal{A}}^{\otimes(k-1)}$, the tensor $\bullet^{\otimes(k-1)} \propto \mathbb{1}$ is the uniform distribution over all contexts.

The incidence, adjacency, and Laplacian operators of G_k can then be expressed as

$$E = \theta_L - \theta_R, \quad A_V = \theta_L \theta_R^* + \theta_R \theta_L^*, \quad \Lambda_V = EE^* = 2I - A_V.$$

3.2 THE VOCABULARY FOURIER TRANSFORM DIAGONALISES THE LAPLACIAN

Following (Philippakis et al., 2024), we can perform a change of basis on $V_{\mathcal{A}}$ whose first vector is the de Bruijn vector $\bullet$, with the remaining orthonormal vectors chosen arbitrarily. This diagonalises the cyclic shift on $\mathcal{A}$. For $q = 2$, the Walsh-Hadamard transform is used; for general q, the discrete Fourier transform. Extending this transform to the context space via $F^{\otimes(k-1)}$ block-diagonalises the Laplacian:

$$\widehat{\Lambda}_V = F^{\otimes(k-1)} \Lambda_V (F^{\otimes(k-1)})^{-1}, \tag{6}$$

with each block corresponding to an independent harmonic mode. As with $\widehat{\Lambda}_V$, we write $\widehat{\theta}_L, \widehat{\theta}_R$ and their adjoints obtained by conjugating θ_L, θ_R (and their adjoints) by the Fourier transform $F^{\otimes(k-1)}$.

Definition 3.3 (Fourier modes)**.** *A* Fourier mode of length i *is any element $\widehat{x} \in \mathcal{X}_{\mathcal{A}}^k[i] \subset V_{\mathcal{A}}^{\otimes i}$ satisfying $\widehat{\theta}_L(\widehat{x}) = \widehat{\theta}_R(\widehat{x}) = 0$, i.e. neither the first nor the last Fourier-basis letter is e_0 (the uniform direction). The set $\mathcal{X}_{\mathcal{A}}^k[i]$ has cardinality $q^{i-2}(q-1)^2$ for $i \geq 2$, $q-1$ for $i = 1$, and 1 for $i = 0$.*

Remark 3.4 (Explicit construction of Fourier modes)**.** *Let $\mathcal{F} = \{f_0, f_1, \ldots, f_{q-1}\}$ be the basis for $V_{\mathcal{A}}$ ($f_0 = \bullet$ being the uniform direction). A basis element of $V_{\mathcal{A}}^{\otimes i}$ has the form $f_{r_0} \otimes \cdots \otimes f_{r_{i-1}}$ with $r_j \in \{0, \ldots, q-1\}$. Such an element belongs to $\mathcal{X}_{\mathcal{A}}^k[i]$ if and only if $r_0 \neq 0$ and $r_{i-1} \neq 0$ (neither the first nor the last index is the uniform direction).*

Remark 3.5. *Each Fourier mode captures a specific pattern of letter correlations among i consecutive tokens, with the constraint that the pattern is not trivially reducible at either end. For $q = 2$ with basis $\{e_0 = \frac{1}{\sqrt{2}}(0 + 1),\ e_1 = \frac{1}{\sqrt{2}}(0 - 1)\}$, length-1 mode e_1 detects single-token bias (preference for 0 or 1), the length-2 mode $e_1 \otimes e_1$ detects a preference for alternation (01 or 10) over runs (00 or 11), while longer modes detect higher-order correlations across more positions.*

3.3 Closed-Form Harmonic Eigenvectors

Each Fourier mode, placed at varying positions within a context, generates a family of eigenvectors via a sine-series construction.

Theorem 3.6 (Harmonic eigenvectors of $\widehat{\Lambda}_V$ (Philippakis et al., 2024, Theorem 3.1))**.** *For each Fourier mode $\widehat{x} \in \mathcal{X}^k_{\mathcal{A}}[i]$ and each $h \in \{1, \ldots, j+1\}$ where $i \geq 1, j = k - 1 - i \geq 0$, the vector*

$$x_h = \sum_{m=1}^{j+1} \sin\Big(\frac{mh\pi}{j+2}\Big)(\widehat{\theta}^*_L)^{j+1-m}(\widehat{\theta}^*_R)^{m-1}\widehat{x} \tag{7}$$

is an eigenvector of $\widehat{\Lambda}_V$ with eigenvalue

$$\lambda_h = 2 - 2\cos\Big(\frac{h\pi}{j+2}\Big). \tag{8}$$

The kernel of $\widehat{\Lambda}_V$ is spanned by $\bullet^{\otimes(k-1)}$ (eigenvalue 0).

Applying the inverse Fourier transform $(F^{\otimes(k-1)})^{-1}$ to each x_h yields an eigenvector of the original vertex Laplacian Λ_V with the same eigenvalue. We refer to both the Fourier-space vectors x_h and their inverse transforms as *harmonic eigenvectors*. For convenience, in later sections we enumerate all non-constant harmonic eigenvectors (eigenvalue > 0) by a single index i, writing $\{x_i\}$.

4 The Harmonic Spectrum as a Fingerprint of Learned Biases

With the harmonic basis in place, we now establish the central claim: the model's stationary distribution decomposes into harmonic modes in exact analogy with a Fourier series, with the uniform walk playing the role of the zero function.

4.1 The Uniform Walk as the Spectral Ground State

When $T \to \infty$, the walk becomes uniform: every transition probability equals $1/q$. This is the baseline against which all learned behaviour is measured (all proofs are in Appendix A).

Proposition 4.1 (Spectral ground state)**.** *For the uniform walk on unweighted G_k, $\Lambda_V = 2I - A_V$ as in equation 3.1, with eigenvectors given by Theorem 3.6 (after inverse Fourier transform). The stationary distribution $\pi^{(\infty)} = q^{-(k-1)}\,\mathbb{1}$ is proportional to $\bullet^{\otimes(k-1)}$, the zero-eigenvalue eigenvector. Its projection onto every non-trivial eigenvector (eigenvalue $\lambda > 0$) is zero.*

This is the analogue of the zero function in classical Fourier analysis.

Corollary 4.2 (Learned biases are harmonic content)**.** *Any deviation of a trained model's stationary distribution $\pi^{(T)}$ from the uniform distribution $\pi^{(\infty)}$ must appear as a non-zero projection onto some Fourier mode x_h with $\lambda_h > 0$.*

4.2 The Temperature-Independent Harmonic Spectrum

We now prove that the projection of the stationary distribution onto each harmonic eigenvector factorises as γ_i/T for large T, where the coefficient γ_i is temperature-independent and determined by the model's centred logits through a Poisson equation. This shows that temperature acts as a simple scaling factor: it does not alter the qualitative structure of the harmonic content, only its amplitude.

Lemma 4.3 (First-order expansion of transition probabilities)**.** *Assume $\max_{s,u} |\ell(s \mid u)| \leq M < \infty$. For every $s \in \mathcal{A}$, $u \in \mathcal{A}^{k-1}$, and all $\beta = 1/T \in [0, 1/(2M)]$,*

$$P_T(s \mid u) = \frac{1}{q} + \beta\,\sigma(s \mid u) + R_2(s, u, \beta), \tag{9}$$

where

$$\sigma(s \mid u) = \frac{\ell(s \mid u) - \bar{\ell}(u)}{q}, \qquad \bar{\ell}(u) = \frac{1}{q}\sum_{s'} \ell(s' \mid u), \tag{10}$$

where $|R_2(s, u, \beta)| \le C\beta^2$ *uniformly in* s *and* u*, with* $C = 4M^2 e$ *and* $\sum_s \sigma(s \mid u) = 0, \forall u$*.*

Remark 4.4 (Linguistic meaning of σ and the matrix S)**.** $\sigma(s \mid u)$ *is the* logit deviation *of token* s *given context* u*: how much the model prefers* s *over the average token. The constraint* $\sum_s \sigma(s \mid u) = 0$ *means that preferences are always relative – there is no "absolute" preference, only contrast.*

We define the $|V| \times |V|$ *perturbation matrix* S *over vertices* $V = \mathcal{A}^{k-1}$ *by*

$$S_{u \to v} = \sigma(v_{k-2} \mid u) \quad \text{if } u_{1:} = v_{:k-2}, \qquad S_{u \to v} = 0 \text{ otherwise}.$$

S *encodes the first-order directional biases of the model as a perturbation to the uniform walk* $T^{(\infty)}$*, so that* $T^{(\beta)} = T^{(\infty)} + \beta S + O(\beta^2)$ *(Lemma 4.3).*

Theorem 4.5 (Temperature-independent harmonic coefficients)**.** *Retain the boundedness assumption of Lemma 4.3. For each non-constant harmonic eigenvector* x_i *of* Λ_V *(the inverse Fourier transform of a Fourier-mode eigenvector from Theorem 3.6) with eigenvalue* $\lambda_i > 0$*, there exists a unique temperature-independent coefficient* $\gamma_i \in \mathbb{R}$*, determined solely by the centred logits, such that*

$$\langle \pi^{(T)} - \pi^{(\infty)}, x_i \rangle = \frac{\gamma_i}{T} + O\left(\frac{1}{T^2}\right) \qquad \text{for all } T \ge 2M. \tag{11}$$

The coefficient γ_i *is given by* $\gamma_i = \langle \pi^{(1)}, \tilde{x}_i \rangle$*, where* $\tilde{x}_i = x_i / \|x_i\|_2$ *is the unit-norm version of the eigenvector* x_i *and where* $\pi^{(1)} \in \mathbb{R}^{|V|}$ *is the unique mean-zero solution of the Poisson equation*

$$(I - (T^{(\infty)})^\top)\pi^{(1)} = S^\top \pi^{(\infty)}, \tag{12}$$

with $T^{(\infty)}$ *the uniform transition matrix (*$T^{(\infty)}_{u \to v} = 1/q$ *for valid transitions) and* $\pi^{(\infty)} = q^{-(k-1)}\mathbf{1}$*. In particular:*

1. γ_i *depends only on the centred logits* $\delta\ell(s \mid u) = \ell(s \mid u) - \bar{\ell}(u)$*; it is independent of temperature.*
2. $\gamma_i = 0$ *for all* i *if and only if, for every vertex* v*, the sum of the logit deviations over incoming contexts vanishes:* $\sum_{u : u_{1:} = v_{:k-2}} \sigma(v_{k-2} \mid u) = 0$*.*

Theorem 4.5 introduces two related quantities. On the one hand, it defines the *temperature-independent harmonic spectrum* as the collection $\{\gamma_i\}_{\lambda_i > 0}$ that characterise the model's learned logit structure. On the other hand, it provides the *observed harmonic projections* at temperature T that are

$$\widehat{\gamma}_i(T) := \langle \pi^{(T)} - \pi^{(\infty)}, x_i \rangle, \qquad \widehat{\Gamma}(T) := \sum_{\lambda_i > 0} \widehat{\gamma}_i(T)^2.$$

By Theorem 4.5, $\widehat{\gamma}_i(T) = \gamma_i / T + O(1/T^2)$ and $\widehat{\Gamma}(T) \to 0$ as $T \to \infty$.

The temperature-independent total harmonic content is defined as $\Gamma := \sum_{\lambda_i > 0} \gamma_i^2$, and it satisfies $\widehat{\Gamma}(T) \approx \Gamma / T^2$ for large T. In the experiments of Section 5, we evaluate at $T = 1$, so the observed projections $\widehat{\gamma}_i(1)$ are close to the intrinsic coefficients γ_i (up to higher-order corrections), and we directly report $\widehat{\gamma}_i(1)$ and $\widehat{\Gamma}(1)$.

Remark 4.6. *Theorem 4.5 is an asymptotic statement, and we do not read the observed quantity* $\widehat{\gamma}_i(1)$ *reported throughout Section 5 as being numerically equal to the* $T \to \infty$ *coefficient* γ_i*. Rather,* $T\,\widehat{\gamma}_i(T)$ *is the natural finite-temperature observable of the same harmonic coordinate system: it coincides with* γ_i *in the limit and, as we verify empirically in Appendix C, varies smoothly with* T *and preserves sign and relative mode structure across a wide temperature range. This makes* $\widehat{\gamma}_i(1)$ *an informative – if not asymptotically exact – proxy for the underlying coefficient, and justifies evaluating at the conventional sampling temperature* $T = 1$ *rather than only in the formal limit.*

Example 4.7 ($k = 3$, $\mathcal{A} = \{0, 1\}$)**.** *The Fourier basis for* $\mathcal{A}$ *is* $e_0 = \frac{1}{\sqrt{2}}(0 + 1)$ *and* $e_1 = \frac{1}{\sqrt{2}}(0 - 1)$*. Ordering contexts as* $(00, 01, 10, 11)$*, the three non-trivial eigenvectors (from Theorem 3.6) are:*

$$\textit{Run asymmetry,} \quad x_1 = \tfrac{1}{\sqrt{2}}(1, 0, 0, -1), \quad \lambda_1 = 1$$

$$\textit{Directional phase,} \quad x_2 = \tfrac{1}{\sqrt{2}}(0, -1, 1, 0), \quad \lambda_2 = 3$$

$$\textit{Alternation bias,} \quad x_3 = \tfrac{1}{2}(1, -1, -1, 1), \quad \lambda_3 = 2$$

Linguistically, $\gamma_1 > 0$ means the model prefers runs of 0*s over runs of* 1*s; $\gamma_2 \neq 0$ captures directional asymmetry in alternation; $\gamma_3 < 0$ means the model prefers switching tokens over repeating them. By Theorem 4.5, the observed projections scale as $\widehat{\gamma}_i(T) = \gamma_i / T$.*

4.3 Numerical Verification of the Spectral Expansion

Theorem 4.5 predicts that $T\,\widehat{\gamma}_i(T) \to \gamma_i$ as $T \to \infty$, with γ_i determined by the Poisson equation equation 12. To verify this, we take $\mathcal{A} = \{0, 1\}$ (so $q = 2$) with context lengths $k = 3$ and $k = 4$ (see Appendix E.2 for more details). For $k = 3$, we let the logit differences be $\Delta\ell(u) = \ell(1 \mid u) - \ell(0 \mid u)$ are chosen as $\Delta\ell(00) = 2, \Delta\ell(01) = -1, \Delta\ell(10) = 0, \Delta\ell(11) = 3$ creating a deliberate asymmetry: the model favours persisting in runs of 1s while making exits from 1s more likely. For each temperature T, we compute the exact stationary distribution $\pi^{(T)}$ from the Kimpton et al. (Kimpton et al., 2022) closed-form expressions and project it onto the non-constant Fourier eigenvectors of Example 4.7.

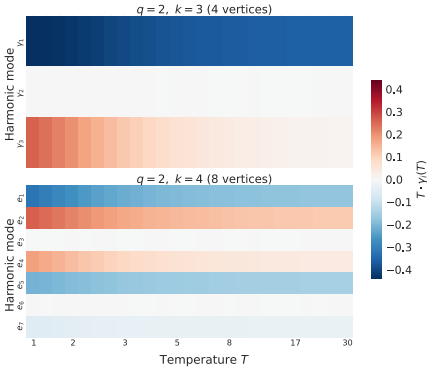

Figure 1: Verification of the perturbation bound for $k = 3$ (top) and $k = 4$ (bottom).

Results. Figure 1 plots $T\,\widehat{\gamma}_i(T)$ as function of T for both $k = 3$ (top panel) and $k = 4$ (bottom panel). The rescaled projections converge rapidly to constant values γ_i as T increases, in precise agreement with the values obtained by solving the Poisson equation equation 12. For $k = 3$, the dominant coefficient is $\gamma_1 \approx -0.354$ (run asymmetry), while $\gamma_3 = 0$ (no net alternation bias for this particular logit choice). For $k = 4$, additional modes become active, and their Poisson-equation predictions are all confirmed. The rapid convergence confirms that the harmonic spectrum is a practically accessible summary of a model's biases, not merely an asymptotic concept. We validate the same observation empirically on real LLMs in Appendix C.

5 Experiments

Theorem 4.5 shows that a model's sequential biases decompose into independent harmonic modes, each with a temperature-invariant coefficient γ_i and each corresponding to a named pattern of token co-occurrence. We probe this decomposition in two settings studied in prior work: binary sequence generation and Rock-Paper-Scissors. Implementation details are in Appendix E.1.

5.1 Experiment 1: Binary random sequence generation

Motivation. Random binary sequence generation is the canonical setting in which non-trivial sequential biases such as the gambler's fallacy – the belief that a run of one outcome makes the opposite outcome more likely – are studied in both humans (Kahneman & Tversky, 1972) and machines. Bigelow et al. (2024) showed that LLMs exhibit exactly this class of bias: their in-context learning of binary sequences is governed by interpretable sequential statistics (run length, alternation, directional preference) that a full next-token distribution obscures, and that persist well beyond what the prompt warrants. To probe the harmonic spectrum more broadly, we introduce two additional deterministic patterns (one-biased, zero-biased) and replace abstract tokens $\{0, 1\}$ with gender tokens $\{\texttt{men}, \texttt{women}\}$. The two leading harmonic coordinates in this regime, run asymmetry $\widehat{\gamma}_1$ and alternation bias $\widehat{\gamma}_3$, are the coefficients that quantify these effects, which is what makes the binary probe the cleanest setting in which to ask whether such biases are vocabulary-independent – in particular, whether a semantically loaded pair is processed through the same sequential statistics as an abstract pair. If the harmonic coordinates are statistically equivalent across vocabularies, it would indicate that LLMs process gender tokens through the same low-level sequential statistics as abstract symbols, with gender biases quantifiable as shifts in run asymmetry and alternation preference.

Setup. We evaluate 27 open-weight autoregressive models (Gemma, Qwen, Llama, Falcon, GPT-2; 1B–14B parameters, see (see Appendix E.3 for the complete list) on standard binary and gender vocabularies. For each pattern (alternating, one-biased, zero-biased), we prompt the model with increasing context length L (4 to 100, step 2) by taking the first L symbols of the periodically extended

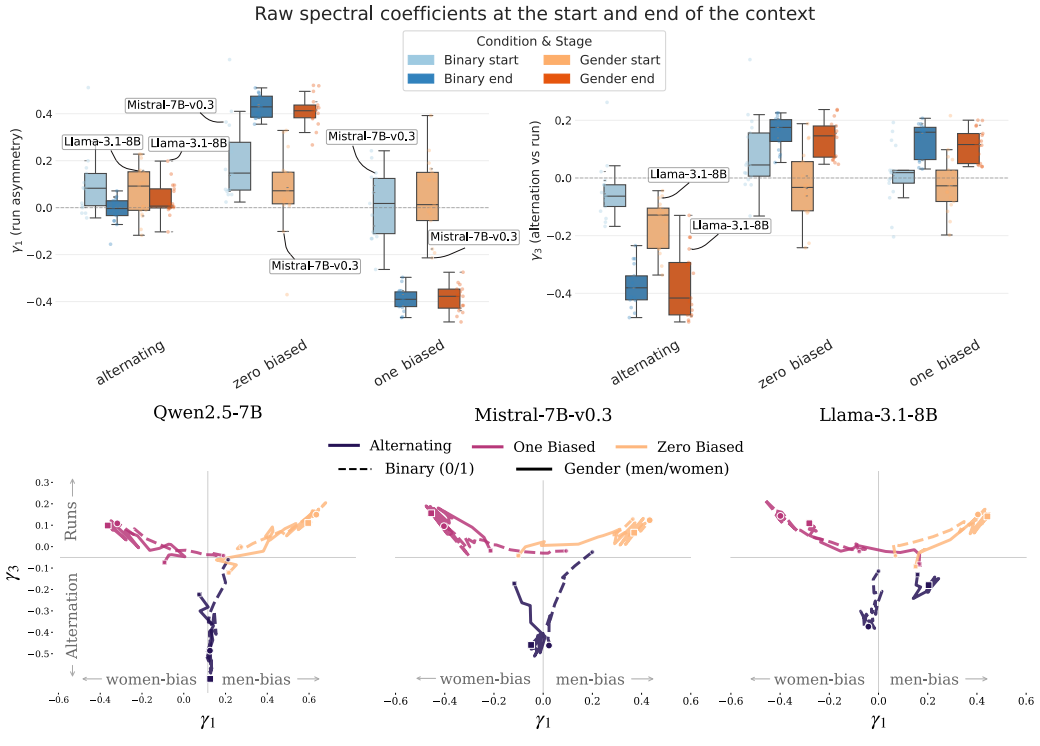

Figure 2: In-context learning trajectories of 17 LLMs in the harmonic spectral plane $(\widehat{\gamma}_1, \widehat{\gamma}_3)$. Top: box plots of start/end points for run asymmetry $\widehat{\gamma}_1$ and alternation bias $\widehat{\gamma}_3$ across three regimes and two vocabularies; Mistral-7B-v0.3 and Llama-3.1-8B are annotated outliers. Bottom: individual ICL paths for Qwen2.5-7B, Mistral-7B-v0.3, and Llama-3.1-8B.

pattern. The prompt uses a simple template inspired by (Bigelow et al., 2024) (e.g., "Generate a sequence of 1000 random choices between men and women…") to avoid extraneous linguistic biases. We retain only the subset of models that reliably follow the instructed pattern (continuation probability $p_{\text{cont}}(L) \geq 0.95$ for all $L \geq 20$, yielding 17 of the 27 models): this filter selects models capable of performing the instructed task at all, since for a model that does not track the pattern, the harmonic coordinates would reflect failure to follow instructions rather than a sequential bias, and the two are not separable without this check.

Results. Figure 2 (top) shows that the three regimes occupy well-separated regions of the $(\widehat{\gamma}_1, \widehat{\gamma}_3)$ plane, and the separation holds across both vocabularies. The binary–gender gap narrows from start to end of context in most regimes, indicating that semantic priors are progressively overridden by in-context evidence; Mistral-7B-v0.3 shows a strong initial `1`/`women` bias that diminishes with longer context, whereas Llama-3.1-8B exhibits a persistent `men`-bias that resists correction, and Qwen2.5-7B shows near-identical binary and gender trajectories, processing gender tokens as abstract symbols throughout. The cross-vocabulary comparison is supported by paired Wilcoxon signed-rank tests (five of six pattern×coordinate comparisons non-significant, the single significant case having effect size 0.024); as non-significance alone does not establish equivalence, we read this as evidence consistent with, rather than proof of, cross-vocabulary equivalence (see the equivalence-margin discussion in Appendix B), and the coordinates are stable under prompt wording and tokenization (sign structure consistent across all 30 pattern×model measurements, see Appendix B).

5.2 Experiment 2: Rock-Paper-Scissors Game

Motivation. Rock-Paper-Scissors (RPS) is a classic paradigm in behavioral game theory for studying systematic, exploitable bias in human sequential decision-making (Wang et al., 2014). In RPS, the unique Nash equilibrium strategy is the uniform mixed strategy that plays R, P, and S each with probability $1/3$ independently of history: it is the only strategy an opponent cannot exploit above the

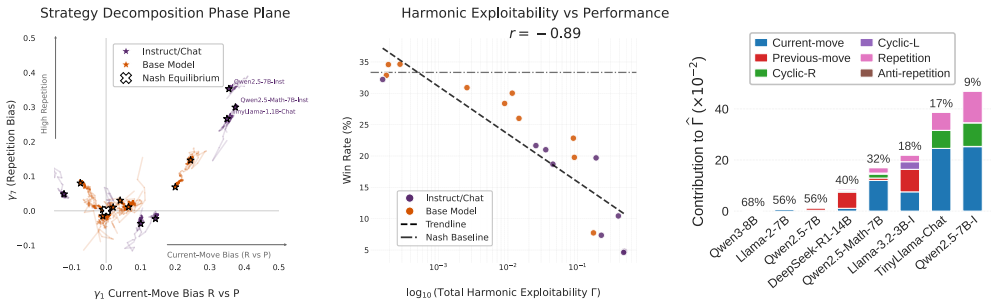

Figure 3: **Harmonic analysis of LLM Rock-Paper-Scissors play**. *Left*: Strategy decomposition phase plane in the $(\widehat{\gamma}_1, \widehat{\gamma}_7)$ subspace. The Nash equilibrium lies at the origin (white $\times$). Instruct/chat models (purple) are consistently further from Nash than base models (orange). *Middle*: Total harmonic departure $\widehat{\Gamma}$ vs. win rate, with Pearson $r = -0.89$ (correlation between $\log_{10} \widehat{\Gamma}$ and win rate). *Right*: Decomposition of $\widehat{\Gamma}$ into its six mode contributions ($\widehat{\gamma}_1^2$, $\widehat{\gamma}_2^2$, $\widehat{\gamma}_5^2$, $\widehat{\gamma}_6^2$, $\widehat{\gamma}_7^2$, $\widehat{\gamma}_8^2$) for eight illustrative models, ordered left-to-right by win+tie rate (annotated above each bar). Full 19-model table in Table 5 in the appendix.

equilibrium payoff, since any systematic deviation from it (e.g., favouring a move, or conditioning on recent history) can in principle be detected and punished. West & Potts (2025) showed that alignment tuning makes LLMs markedly more deterministic and exploitable than their base counterparts in this setting. Our framework therefore offers a spectral explanation for the phenomenon of West & Potts (2025): the total observed harmonic departure $\widehat{\Gamma}(T)$ should directly predict how badly a model will be exploited by a deterministic adversary, and the mode decomposition should reveal which named bias is responsible when two models with similar total bias fail to the same degree.

RPS as a ternary de Bruijn walk. Let $\mathcal{A} = \{R, P, S\}$ ($q = 3$). A model with context length $k-1$ plays RPS by sampling from $P_T(\cdot \mid u)$ where $u \in \mathcal{A}^{k-1}$ is its own recent history. This is exactly a weighted random walk on the ternary de Bruijn graph G_k, and the observed harmonic projections $\{\widehat{\gamma}_i(T)\}$ characterise the model's systematic deviations from Nash. For $k = 3$ (bigram context), the eight non-trivial projections decompose into interpretable components: 1) $\widehat{\gamma}_1, \widehat{\gamma}_2$: current-move bias (preference for R over P/S); 2) $\widehat{\gamma}_3, \widehat{\gamma}_4$: previous-move bias (conditioning on the preceding move); 3) $\widehat{\gamma}_5$ (cyclic-R) / $\widehat{\gamma}_6$ (cyclic-L): tendency to follow or lead the RPS cycle; 4) $\widehat{\gamma}_7$: repetition bias (preference for repeating the last move); 5) $\widehat{\gamma}_8$: anti-repetition bias (preference for switching away). The total observed harmonic departure is $\widehat{\Gamma} := \sum_{i=1}^{8} \widehat{\gamma}_i(T)^2$, evaluated at $T = 1$. As we show in Appendix E.4, $\widehat{\Gamma}$ is exactly proportional to the Pearson χ^2 statistic for departure from the uniform bigram distribution, so the scalar $\widehat{\Gamma}$ alone carries the aggregate information of a standard non-uniformity test; the value the harmonic basis adds is the *mode-wise* decomposition into named components below, which a single aggregate statistic cannot provide.

Setup. We evaluate 19 open-weight language models (both base and instruct-tuned variants) from the Gemma, Qwen, Llama, Falcon, and GPT-2 families. Each model plays 500 independent games, each consisting of 10 rounds, against a *greedy adversary* that knows the model's current context u (the last two moves played) and the corresponding next-token distribution $P_T(\cdot \mid u)$, and always plays the move that beats the model's most probable response (with arbitrary tie-breaking). The model's move probabilities are obtained by applying softmax to the raw logits at temperature $T = 1.0$ and renormalising the probabilities of the single tokens `R`, `P`, `S` to sum to one. Transition probabilities are estimated over the nine bigram contexts, computed by averaging the model's next-token softmax probabilities over all observed bigram contexts. The stationary distribution $\widehat{\pi}$ is computed exactly. Win rates are measured against the same greedy adversary.

Results. Figure 3 (left) plots each model in the $(\widehat{\gamma}_1, \widehat{\gamma}_7)$ plane – current-move bias against repetition bias. Two clusters emerge far from Nash: (1) *positive* $\widehat{\gamma}_1$ + *positive* $\widehat{\gamma}_7$ (upper right): Qwen2.5-7B-Inst, Qwen2.5-Math-7B-Inst, and TinyLlama-1.1B-Chat combine current-move bias with repetition bias – concretely, a preference for one move *and* a tendency to repeat whatever was just played. This is the easiest pattern an adversary can exploit: knowing only that the model favours, say, R and tends to repeat its last move, the adversary can play P immediately after every R and win far more than one third of the time, without needing to model anything else about the policy. Unsurprisingly, these three

models are trivially exploited; (2) *near-Nash* (origin): Gemma-3-1B-pt, Gemma-3-4B-it, Qwen3-1.7B, and Qwen3-8B sit at $\widehat{\Gamma} \approx 0$ with near-uniform play – they are close to genuinely unpredictable and cannot be beaten above the 33.3% baseline by any fixed adversary strategy. Instruct-tuned models are on average further from Nash than base models, consistent with West & Potts (2025).

Figure 3 (middle) shows $\widehat{\Gamma}$ predicts win rate at $r = -0.89$; models near Nash achieve win rates near the 33.3% baseline, while those with large $\widehat{\Gamma}$ are exploited to single digits. As noted above, this aggregate correlation is expected given the exact χ^2-equivalence of Appendix E.4: a large $\widehat{\Gamma}$ simply says that a model's play is far from uniform, and any systematic, non-uniform strategy is in principle punishable by an adversary that has already observed it. The value the harmonic basis adds over this single number is the mode-wise decomposition (Figure 3, right), which separates models with comparable $\widehat{\Gamma}$ but different win rates by showing *which* bias is responsible. The most exploited models are dominated by current-move and repetition bias – exactly the two patterns that, as the example above illustrates. DeepSeek-R1-Qwen-14B is the instructive counterexample: despite a moderate $\widehat{\Gamma}$, its harmonic mass is concentrated almost entirely in previous-move bias – a preference that shows up only once the model's own preceding move is conditioned on, and is therefore invisible to a plain count of how often it plays R, P, or S. Cyclic bias matters only for the most collapsed models, and anti-repetition bias is negligible throughout, suggesting that whatever produces move-to-move dependence in these models, actively avoiding repetition is not part of it.

6 Related Work

Spectral theory of de Bruijn graphs. The eigenvalues of the de Bruijn graph Laplacian were computed by Strok (Strok, 1992) and Delorme and Tillich (Delorme & Tillich, 1998). The closed-form eigenvectors and their connection to cycle and cut spaces are due to Philippakis et al. (Philippakis et al., 2024), whose results form the technical backbone of this work.

De Bruijn graphs in machine learning. n-gram language models are long understood as empirical distributions over the edges of a de Bruijn graph (Jurafsky & Martin, 2009), and de Bruijn graphs are foundational to genome assembly (Compeau et al., 2011; Vrček et al., 2022; Kapuśniak et al., 2023; Wei et al., 2024), providing a rich algorithmic toolkit for sequence analysis. (Qarkaxhija et al., 2022) generalise graph neural network message passing to higher-order de Bruijn graphs, learning causality-aware representations of walks on dynamic, time-stamped graphs. (Saifuddin et al., 2023) used de Bruin graphs for general sequence classification via hypergraphs. As discussed in Section 1, none of these lines of work treats a language model's own autoregressive inference as a walk on the de Bruijn graph and builds an auditing coordinate system from its Laplacian spectrum.

Physics of language models. Recent work studies transformers on synthetic formal languages to understand what structures are learnable and how (Allen-Zhu & Li, 2023; Bhattamishra et al., 2020). Our framework is complementary: rather than asking what structures transformers can express, we ask how the learned distribution is *organised spectrally*.

7 Limitations and Future Work

We have developed a harmonic analysis of autoregressive language model inference, grounding the framework in the spectral theory of de Bruijn graphs. The harmonic spectrum provides a compact, temperature-independent fingerprint of a model's sequential biases, decomposing them into interpretable Fourier modes that range from simple run/alternation preferences to high-order oscillatory patterns. Our empirical evaluation spans 25 different LLMs and two distinct experimental setups, each involving substantial independent validation.

While harmonic framework is exact for any context order k and vocabulary size q, its direct application requires the state-space size q^{k-1} to be manageable. For this, quantisation steps such as part-of-speech tagging (collapsing a large lexicon to $\sim$15 syntactic tags or mapping to dialogue acts) can be used. Our part-of-speech experiment (Appendix D) is a first feasibility demonstration of this route. Extending the framework to larger (k, q) through sparse estimation or tensor decompositions, and applying it to these quantised representations, are natural directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools to generate synthetic data sets, assist with translation, clean and reformat dataset. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, summarize or analyse existing literature, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references, propose a title. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, LLM-generated code was verified and tested for correctness by the authors, mathematical proofs were proofread. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We submit the code to reproduce the results of this paper as supplementary material.

REFERENCES

Z. Allen-Zhu and Y. Li. Physics of language models: Part 1, context-free grammar. arXiv:2305.13673, 2023.

S. Bhattamishra, K. Ahuja, and N. Goyal. On the ability and limitations of transformers to recognize formal languages. In *EMNLP*, 2020.

E. J. Bigelow, E. S. Lubana, R. P. Dick, H. Tanaka, and T. D. Ullman. In-context learning dynamics with random binary sequences. In *ICLR*, 2024.

P. E. C. Compeau, P. A. Pevzner, and G. Tesler. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29:987–991, 2011.

N. G. de Bruijn. A combinatorial problem. *Indagationes Mathematicae*, 8:461–467, 1946.

C. Delorme and J.-P. Tillich. The spectrum of de Bruijn and kautz graphs. *European Journal of Combinatorics*, 19:307–319, 1998.

I. J. Good. Normal recurring decimals. *Journal of the London Mathematical Society*, 21:167–169, 1946.

D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice Hall, 2nd edition, 2009.

Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454, 1972.

Kacper Kapuśniak, Manuel Burger, Gunnar Rätsch, and Amir Joudaki. Learning genomic sequence representations using graph neural networks over de Bruijn graphs. *arXiv preprint arXiv:2312.03865*, 2023.

L. Kimpton, P. Challenor, and H. Wynn. Modelling correlated bernoulli data part i: Theory and run lengths. arXiv:2211.16921 [stat.ME], 2022.

A. Philippakis, N. Mallinar, P. Pandit, and M. Belkin. Eigenvectors of the de Bruijn graph laplacian: A natural basis for the cut and cycle space. arXiv:2410.07622 [math.CO], 2024.

Lisi Qarkaxhija, Vincenzo Perri, and Ingo Scholtes. De Bruijn goes neural: Causality-aware graph neural networks for time series data on dynamic graphs. In *Proceedings of the First Learning on Graphs Conference*, volume 198 of *Proceedings of Machine Learning Research*, pp. 51:1–51:21. PMLR, 2022.

Khaled Mohammed Saifuddin, Corey May, Farhan Tanvir, Muhammad Ifte Khairul Islam, and Esra Akbas. Seq-HyGAN: Sequence classification via hypergraph attention network. In *Proceedings of the 19th International Workshop on Mining and Learning with Graphs (MLG '23)*, 2023.

V. V. Strok. Circulant matrices and the spectra of de Bruijn graphs. *Ukrainian Mathematical Journal*, 44:1446–1454, 1992.

Lovro Vrček, Xavier Bresson, Thomas Laurent, Martin Schmitz, and Mile Šikić. Learning to untangle genome assembly with graph convolutional networks. *arXiv preprint arXiv:2206.00668*, 2022.

Zhijian Wang, Bin Xu, and Hai-Jun Zhou. Social cycling and conditional responses in the Rock-Paper-Scissors game. *Scientific Reports*, 4:5830, 2014. doi: 10.1038/srep05830.

Yuxiao Wei, Qi Zhang, and Liwei Liu. The improved de Bruijn graph for multitask learning: predicting functions, subcellular localization, and interactions of noncoding RNAs. *Briefings in Bioinformatics*, 26(1):bbae627, 2024. doi: 10.1093/bib/bbae627.

P. West and C. Potts. Base models beat aligned models at randomness and creativity. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025.

O. Zekri, A. Odonnat, A. Benechehab, L. Bleistein, N. Boullé, and I. Redko. Large language models as Markov chains. arXiv preprint arXiv:2410.02724, 2024.

A Proofs

A.1 Proof of Proposition 4.1

Proof. G_k is q-regular, so the uniform walk is doubly stochastic. Its unique stationary distribution is the uniform vector. Since $\bullet^{\otimes(k-1)}$ spans $\ker(\Lambda_V)$ and $\pi^{(\infty)} \propto \bullet^{\otimes(k-1)}$, orthogonality of eigenvectors gives $\langle \pi^{(\infty)}, x_h \rangle = 0$ for all $\lambda_h > 0$. □

A.2 Proof of Lemma 4.3

Write $\delta\ell(s \mid u) := \ell(s \mid u) - \bar{\ell}(u)$, so $\sum_{s'} \delta\ell(s' \mid u) = 0$ and $|\delta\ell| \le 2M$. Since $e^{-\beta\bar{\ell}(u)}$ cancels between numerator and denominator,

$$P_T(s \mid u) = \frac{e^{\beta\delta\ell(s|u)}}{\sum_{s'} e^{\beta\delta\ell(s'|u)}}.$$

Applying the Taylor remainder theorem to $t \mapsto e^t$ on $|t| \le 2\beta M$ gives numerator $= 1 + \beta\delta\ell + r_1$ with $|r_1| \le 2\beta^2 M^2 e^{2\beta M}$. Summing over s' and using $\sum \delta\ell = 0$ gives denominator $= q + \sum r_1$, and $|q^{-1} - Z(\beta, u)^{-1}| = O(\beta^2)$ for $\beta \le 1/(2M)$. Multiplying and collecting gives equation 9 with $\sigma = \delta\ell/q$.

A.3 Proof of Theorem 4.5

Step 1. By Lemma 4.3, write $T^{(\beta)} = T^{(\infty)} + \beta S + E(\beta)$ with $\|E(\beta)\|_\infty \le C\beta^2$.

Step 2 (existence and uniqueness of $\pi^{(1)}$). $I - (T^{(\infty)})^\top$ has kernel $\operatorname{span}(\mathbb{1})$ (since $T^{(\infty)}$ is doubly stochastic and G_k is strongly connected). The right-hand side $b = S^\top \pi^{(\infty)} \propto S^\top \mathbb{1}$ satisfies $\mathbb{1}^\top b = 0$ because $S\mathbb{1} = 0$. By the Fredholm alternative, equation 12 has a solution; the mean-zero constraint $\langle \pi^{(1)}, \mathbb{1} \rangle = 0$ selects it uniquely.

Step 3 (second-order expansion). Write $\pi^{(\beta)} = \pi^{(\infty)} + \beta\pi^{(1)} + O(\beta^2)$ and substitute into $(T^{(\beta)})^\top \pi^{(\beta)} = \pi^{(\beta)}$. The order-$\beta^0$ equation is satisfied by $\pi^{(\infty)}$; the order-β^1 equation is exactly equation 12. Analyticity of $\beta \mapsto \pi^{(\beta)}$ (by the analytic implicit function theorem applied to the stationarity equation, using that the spectral gap of $T^{(\infty)}$ is 1) gives $\|\pi^{(\beta)} - \pi^{(\infty)} - \beta\pi^{(1)}\|_1 \le K\beta^2$ for explicit K depending only on M, q, and k.

Step 4 (projection). Let $\tilde{x}_i = x_i/\|x_i\|_2$ where, by the discrete Parseval identity,

$$\|x_i\|_2^2 \;=\; \sum_{m=1}^{j+1} \sin^2\!\left(\frac{mh\pi}{j+2}\right) \;=\; \frac{j+2}{2},$$

so $\|\tilde{x}_i\|_2 = 1$. Since $\langle \pi^{(\infty)}, \tilde{x}_i \rangle = 0$ (Proposition 4.1) and $\|\tilde{x}_i\|_2 = 1$,

$$\langle \pi^{(\beta)}, \tilde{x}_i \rangle \;=\; \beta \langle \pi^{(1)}, \tilde{x}_i \rangle + O(\beta^2).$$

Substituting $\beta = 1/T$ and $\gamma_i := \langle \pi^{(1)}, \tilde{x}_i \rangle$ yields equation 11.

Items (1) and (2) follow from Step 2: $\pi^{(1)}$ solves the linear system equation 12 whose right-hand side depends only on S (hence only on $\delta\ell$), establishing (1); and $\pi^{(1)} = 0$ iff $S^\top \mathbb{1} = 0$, which is the stated column-sum condition, establishing (2).

B Statistical Robustness of the Binary/Gender Vocabulary Comparison

This appendix gives the detailed statistical support for the cross-vocabulary claim in Section 5.1, together with a separate check that the harmonic coordinates for a fixed vocabulary are themselves stable under superficial prompt changes.

Statistical support for binary/gender equivalence. We test the visual overlap in Figure 2 formally with paired Wilcoxon signed-rank tests, per pattern, across the $N = 17$ reliable models (those with $p_{\text{cont}} \geq 0.95$), pairing each model's binary- and gender-vocabulary harmonic coordinates. Five of the six pattern×coordinate comparisons are not significant; only the zero-biased $\Delta\gamma_3$ comparison reaches $p < 0.05$, and with a small effect size (0.157 vs. 0.133, a difference of 0.024). Representative close matches include the one-biased $\Delta\gamma_1$ coordinate, -0.389 ± 0.048 (binary) vs. -0.387 ± 0.056 (gender), and the zero-biased $\Delta\gamma_1$ coordinate, 0.427 ± 0.047 vs. 0.413 ± 0.065. We stress that a non-significant paired test does not by itself establish equivalence (absence of evidence is not evidence of absence); with $N = 17$ the tests have limited power to detect small but real differences. We accordingly read the cross-vocabulary claim conservatively: binary and gender realisations are statistically indistinguishable in the large majority of tested harmonic coordinates, with at most a weak effect in one coordinate, and we would welcome a pre-registered equivalence test (e.g. TOST with a stated margin) on a larger model set as a stronger follow-up than the significance tests reported here.

Robustness to prompt wording and tokenization. The equivalence above concerns two specific vocabularies; a separate question is whether the harmonic coordinates for a fixed vocabulary are themselves stable under superficial prompt changes. We check this with two smaller open models (Qwen3-0.6B, Qwen3-1.7B) across the three deterministic patterns and five surface variants at fixed context length $L = 24$: `digits_instruction` and `digits_qa` (both `0/1`, different wording), `yes_no`, `heads_tails`, and `men_women` (same instruction wording, different labels). The underlying bit process is identical across variants; only surface rendering changes. Table 1 reports (γ_1, γ_3) and the continuation probability, as mean $\pm$ std across the five variants. Across all 30 pattern×model measurements, the sign structure is perfectly consistent: alternating always yields $\gamma_3 < 0$, one-biased always yields $\gamma_1 < 0, \gamma_3 > 0$, and zero-biased always yields $\gamma_1 > 0, \gamma_3 > 0$, with continuation probability remaining high throughout. This supports a conservative robustness statement: the harmonic descriptor is stable under controlled wording and tokenization changes, and is not an artifact of one particular surface form.

C Finite-Temperature Behaviour of Real Models

Theorem 4.5 and its verification in Section 4.3 concern the closed-form toy walk, where $T\,\widehat{\gamma}_i(T)$ converges to the Poisson-equation solution γ_i exactly. We complement this with an empirical check on real LLMs, addressing the concern that the finite-temperature quantities reported in Section 5 (evaluated at the conventional sampling temperature $T = 1$) might not be representative of the asymptotic coefficients they are meant to approximate.

Table 1: Robustness of the harmonic coordinates to prompt wording and label choice. Each row reports mean ± std across five surface variants (`digits_instruction`, `digits_qa`, `yes_no`, `heads_tails`, `men_women`) of the identical underlying binary process, at $L = 24$.

Model	Pattern	γ_1 (mean ± std)	γ_3 (mean ± std)	cont. prob. (mean ± std)
Qwen3-0.6B	alternating	0.0052 ± 0.0068	-0.4854 ± 0.0137	0.9925 ± 0.0071
Qwen3-0.6B	one-biased	-0.4532 ± 0.0150	0.1562 ± 0.0101	0.9620 ± 0.0279
Qwen3-0.6B	zero-biased	0.4466 ± 0.0326	0.1378 ± 0.0490	0.9375 ± 0.0295
Qwen3-1.7B	alternating	-0.00001 ± 0.00004	-0.49991 ± 0.00007	0.99996 ± 0.00003
Qwen3-1.7B	one-biased	-0.4696 ± 0.0033	0.1651 ± 0.0048	0.9939 ± 0.0030
Qwen3-1.7B	zero-biased	0.4815 ± 0.0297	0.1837 ± 0.0397	0.9786 ± 0.0223

For each of three models (TinyLlama-1.1B-Chat, Qwen3-1.7B, Qwen2.5-1.5B-Instruct), we compute the dominant binary harmonic coordinate $T\,\widehat{\gamma}_1(T)$ across a temperature sweep $T \in \{0.5, 1.0, 2.0, 3.0, 5.0, 8.0, 12.0, 20.0\}$ and compare it to the $T \to \infty$ coefficient γ_1 obtained from the Poisson equation equation 12, using the alternating-pattern binary probe of Experiment 1 at context length $L = 30$.

Table 2: Finite-temperature dominant harmonic coordinate $T\,\widehat{\gamma}_1(T)$ compared to the $T \to \infty$ Poisson-equation coefficient γ_1.

Model	$T = 0.5$	$T = 1.0$	$T = 2.0$	$T = 3.0$	$T = 5.0$	$T = 8.0$	$T = 12.0$	$T = 20.0$	γ_1
TinyLlama-1.1B-Chat	+0.3314	+0.4811	+0.5717	+0.5868	+0.5892	+0.5849	+0.5810	+0.5772	+0.5702
Qwen3-1.7B	+0.3662	+0.6993	+1.2191	+1.4212	+1.4997	+1.4705	+1.4320	+1.3943	+1.3247
Qwen2.5-1.5B-Instruct	+0.3530	+0.5604	+0.7217	+0.7593	+0.7733	+0.7711	+0.7665	+0.7615	+0.7511

The observed coordinate moves smoothly and monotonically from $T = 0.5$ toward the $T \to \infty$ target, overshooting slightly around $T \approx 5$–8 before relaxing, but never changing sign or crossing to a qualitatively different regime. The same behaviour holds for the secondary coordinate: for Qwen3-1.7B at $L = 30$, $T\widehat{\gamma}_1(T)$ moves from -0.6486 at $T = 1$ to -0.8820 at the Poisson target, while $T\widehat{\gamma}_3(T)$ moves from $+0.4246$ to $+0.2395$ over the same range, again without a sign change. This supports Remark 4.6: the value reported at $T = 1$ throughout the paper is not the asymptotic coefficient itself, but it is a stable, sign-preserving, and monotonically-informative proxy for it, rather than an artifact specific to one temperature choice.

D Feasibility Extension: Open-Ended Generation via Part-of-Speech Sequences

The experiments in the main text (binary sequences, RPS) restrict the model's output to a small, closed alphabet, which is what makes the induced walk observable at all. As a feasibility check rather than a core result, we asked whether the harmonic framework also says something useful about ordinary, open-ended generation, by lifting a genuinely linguistic signal – the part-of-speech (POS) tag of each generated word – onto a $q = 12$ Universal POS (UPOS) alphabet at de Bruijn order $k = 3$ (bigram context).

Setup. The reference stream is the first 40,000 tokens of the Brown corpus (`brown.words()[:40000]`). Generated text is sampled from Qwen2.5-0.5B-Instruct using the prompt "Write a long fluent English article about {topic}. At least 400 words of prose." over a rotating topic list, at temperatures $T \in \{0.7, 1.0, 1.3\}$, collecting approximately 15,000 valid UPOS tags per temperature. Both streams are tagged with the same UPOS tagger, and the POS bigram transition operator is estimated and projected onto the de Bruijn harmonic eigenbasis of Theorem 3.6 for $q = 12$, $k = 3$. As a null control for finite-sample artifacts, each tag sequence is compared to a shuffled version of itself, which preserves unigram tag frequencies but destroys their order. We report the total harmonic mass $\Gamma(T) = \sum_{i:\lambda_i>0} \gamma_i(T)^2$ and the ordering-specific quantity

$$\Delta\Gamma(T) = \Gamma(T; \text{POS sequence}) - \Gamma(T; \text{shuffled control}).$$

Results. Table 3 shows that ordering-specific harmonic mass is large and clearly non-zero for both reference English and model-generated text against its own shuffled null, indicating that the harmonic

Table 3: Harmonic mass on a 12-symbol UPOS alphabet ($k = 3$). $\Delta\Gamma(T)$ isolates ordering-specific structure against a frequency-matched shuffled control.

Regime	n (tags)	$\Delta\Gamma(T)$
Reference English	35,401	0.121
Qwen-0.5B-Inst, $T = 0.7$	15,077	0.098
Qwen-0.5B-Inst, $T = 1.0$	15,048	0.086
Qwen-0.5B-Inst, $T = 1.3$	15,262	0.082

decomposition picks up real sequential POS structure on a larger, linguistic alphabet rather than only on 2–3 symbol toy vocabularies. $\Delta\Gamma(T)$ also decreases smoothly and monotonically as temperature rises from 0.7 to 1.3, consistent with the model's syntactic structure being progressively randomised at higher temperature. We had also computed the cosine alignment between each model's harmonic coefficient vector and the reference-English coefficient vector; we do not report it here, since without an appropriate null (e.g., shuffled-vs-reference, or independent-sample-of-English baselines) a single cosine number is not readily interpretable, and we would rather understate this feasibility check than overstate it. We treat this experiment as a proof of concept that the framework extends to open-ended generation, not as a full scaling study: a broader model sweep, longer context orders k, and a principled baseline for the alignment-to-reference question are left to future work.

E Experimental details

E.1 Practical Estimation of the Harmonic Coefficients

Theorem 4.5 provides a precise characterisation of the temperature-independent coefficients γ_i: they are the projections $\langle \pi^{(1)}, x_i \rangle$ of the solution $\pi^{(1)}$ to the Poisson equation

$$(I - (T^{(\infty)})^\top)\pi^{(1)} = S^\top \pi^{(\infty)}, \tag{13}$$

where the matrix S encodes the centred logit deviations $S_{u\to v} = \frac{1}{q}\delta\ell(v_{k-2} \mid u)$ for valid transitions. In principle, given white-box access to the logits $\ell(s \mid u)$, one could compute S, solve equation 13 for $\pi^{(1)}$, and project onto the Fourier modes x_i.

In practice, however, one rarely has direct access to the raw logits; instead, one observes the model's behaviour through its transition probabilities $P_T(s \mid u)$ at a finite temperature T. The theorem shows that for large T,

$$\widehat{\gamma}_i(T) = \langle \pi^{(T)} - \pi^{(\infty)}, x_i \rangle \approx \frac{\gamma_i}{T},$$

so that the desired coefficient is asymptotically $\gamma_i \approx T\,\widehat{\gamma}_i(T)$. The following procedure, which mirrors the experimental pipeline used in this work, estimates the harmonic spectrum from finite-temperature samples.

1. **Estimate the transition matrix $\widehat{T}$.** For a given pattern of length L (e.g., an alternating binary sequence), iterate over all cyclic shifts of the pattern to remove phase-alignment artefacts. For each shift, and for each $(k-1)$-gram context u occurring in the shifted sequence, query the model to obtain the next-token distribution $P_T(\cdot \mid u)$. Accumulate these probabilities in a matrix $\widehat{T} \in \mathbb{R}^{q^{k-1}\times q}$, where $\widehat{T}[u, s]$ is the average probability of token s following context u.

2. **Compute the stationary distribution $\widehat{\pi^{(T)}}$.** From $\widehat{T}$, construct the full $q^{k-1} \times q^{k-1}$ transition matrix $\widehat{P}$ using the de Bruijn overlap condition (Equation (2) in the main text):

   $$\widehat{P}_{u\to v} = \widehat{T}[u,\, v_{k-2}] \quad \text{if } u_{1:} = v_{:k-2},$$

   and zero otherwise. The stationary distribution $\widehat{\pi^{(T)}}$ is the unique probability vector satisfying $\widehat{\pi^{(T)}}\,\widehat{P} = \widehat{\pi^{(T)}}$. In practice this is solved by appending the normalisation constraint $\sum_v \widehat{\pi^{(T)}}(v) = 1$ and using a least-squares solver.

3. **Project onto the Fourier basis.** For each non-constant eigenvector x_i (given analytically by Theorem 3.6), compute the finite-temperature projection

$$\widehat{\widehat{\gamma}}_i(T) = \langle \widehat{\pi^{(T)}} - \pi^{(\infty)}, x_i \rangle.$$

(Notice the double hat only denotes the estimation from finite samples; in the analysis we treat it as $\widehat{\gamma}_i(T)$.)

4. **Extrapolate to temperature-independent coefficients.** By Theorem 4.5, $\widehat{\gamma}_i(T) \approx \gamma_i / T$. Therefore, set

$$\widehat{\gamma}_i = T \cdot \widehat{\gamma}_i(T).$$

Note that for $q = 2$ the DFT reduces to the Walsh–Hadamard transform, which is real-valued, so the eigenvectors of Theorem 3.5 are directly real. For $q > 2$ the DFT character associated with a non-trivial pair (k_a, k_b) is the complex vector

$$\chi_{k_a,k_b}(a,b) = \omega^{k_a a + k_b b}, \qquad \omega = e^{2\pi i/q}, \quad (a,b) \in \mathcal{A}^2,$$

and its conjugate $\overline{\chi}_{k_a,k_b} = \chi_{q-k_a,\, q-k_b}$ is a distinct non-trivial character (since $(k_a, k_b) \neq (0,0)$ and $q > 2$). The two characters span a two-dimensional eigenspace of Λ_V with the same eigenvalue. Since Λ_V is real and symmetric, this eigenspace has a real basis, given explicitly by the normalised real and imaginary parts:

$$\tilde{x}^{\mathrm{Re}}_{k_a,k_b} = \frac{\mathrm{Re}(\chi_{k_a,k_b})}{\|\mathrm{Re}(\chi_{k_a,k_b})\|_2}, \qquad \tilde{x}^{\mathrm{Im}}_{k_a,k_b} = \frac{\mathrm{Im}(\chi_{k_a,k_b})}{\|\mathrm{Im}(\chi_{k_a,k_b})\|_2}.$$

These two vectors are orthogonal (since $\langle \mathrm{Re}(\chi), \mathrm{Im}(\chi) \rangle = \frac{1}{2} \mathrm{Im}\langle \chi, \chi \rangle = 0$ for a DFT character with $\chi \neq \bar{\chi}$), unit-norm by construction, and together span the same eigenspace as χ_{k_a,k_b} and $\overline{\chi}_{k_a,k_b}$. Collecting these pairs over all non-trivial (k_a, k_b) up to conjugacy yields a complete real orthonormal eigenbasis $\{\tilde{x}_i\}$ of Λ_V. The harmonic coefficients $\gamma_i = \langle \pi^{(1)}, \tilde{x}_i \rangle$ are inner products of two real vectors and are therefore real for all q, with no complex parts to discard. For the ternary RPS experiment ($q = 3$, $k = 3$) this produces exactly eight real eigenvectors, corresponding to the four DFT pairs $(k_a, k_b) \in \{(0,1), (1,0), (1,1), (1,2)\}$ and their conjugates; the resulting $\gamma_1, \ldots, \gamma_8$ are the coefficients reported in Table 1.

Connection to the Poisson equation. The matrix S in equation 13 is the first-order term in the expansion $P_T(s \mid u) = 1/q + \frac{1}{T}\sigma(s \mid u) + O(1/T^2)$ (Lemma 4.3), with $\sigma(s \mid u) = \delta\ell(s \mid u)/q$. The estimated transition matrix $\widehat{T}$ at temperature T therefore approximates $T^{(\beta)} = T^{(\infty)} + \beta S + O(\beta^2)$ with $\beta = 1/T$. The stationary distribution $\widehat{\pi^{(T)}}$ computed from $\widehat{T}$ satisfies $\widehat{\pi^{(T)}} = \pi^{(\infty)} + \beta\pi^{(1)} + O(\beta^2)$, where $\pi^{(1)}$ is the solution to equation 13. Consequently,

$$\langle \widehat{\pi^{(T)}} - \pi^{(\infty)}, x_i \rangle = \beta \langle \pi^{(1)}, x_i \rangle + O(\beta^2) = \frac{\gamma_i}{T} + O(1/T^2),$$

justifying the extrapolation $\widehat{\gamma}_i = T\,\widehat{\gamma}_i(T)$.

In the binary case $q = 2$, $k = 3$, the eigenvectors $\{x_1, x_2, x_3\}$ are given explicitly in Example 4.7. The structural symmetry of the binary de Bruijn graph forces $\widehat{\gamma}_2(T) = 0$ identically, so only $\widehat{\gamma}_1$ (run asymmetry) and $\widehat{\gamma}_3$ (alternation bias) need to be estimated.

E.2 Logit values for Section 4.3

Logit specification. We take $\mathcal{A} = \{0, 1\}$ ($q = 2$) and verify at context lengths $k = 3$ and $k = 4$. For $k = 3$ the model is parameterised by four logit differences $\Delta\ell(u) = \ell(1 \mid u) - \ell(0 \mid u)$, one per bigram context $u \in \{00, 01, 10, 11\}$:

$$\Delta\ell(00) = 2.0, \quad \Delta\ell(01) = -1.0, \quad \Delta\ell(10) = 0.0, \quad \Delta\ell(11) = 3.0.$$

For $k = 4$ the model is parameterised by eight logit differences, one per trigram context $u \in \{000, 001, 010, 011, 100, 101, 110, 111\}$:

$$\Delta\ell(000) = 2.0, \quad \Delta\ell(001) = -1.0, \quad \Delta\ell(010) = 0.5, \quad \Delta\ell(011) = 1.5,$$
$$\Delta\ell(100) = -0.5, \quad \Delta\ell(101) = 1.0, \quad \Delta\ell(110) = -1.5, \quad \Delta\ell(111) = 2.5.$$

Table 4: All 27 open-weight autoregressive language models evaluated in the binary sequence experiment, listed by family. Models with $p_{\text{cont}}(L) \geq 0.95$ for all $L \geq 20$ (17 models) are used in Figure 2; the remainder are excluded because they do not reliably follow the instructed pattern, so their harmonic coordinates would not be separable from instruction-following failure.

Family	Model ID (Hugging Face)

Gemma	`google/gemma-3-1b-pt`
	`google/gemma-3-1b-it`
	`google/gemma-3-4b-it`
	`google/gemma-3-4b-pt`
	`google/gemma-2-9b-pt`
	`google/gemma-2-9b-it`

Qwen	`Qwen/Qwen2.5-7B`
	`Qwen/Qwen2.5-7B-Instruct`
	`Qwen/Qwen2.5-Math-7B`
	`Qwen/Qwen2.5-Math-7B-Instruct`
	`Qwen/Qwen2.5-Coder-7B-Instruct`
	`Qwen/Qwen3-1.7B`
	`Qwen/Qwen3-4B`
	`Qwen/Qwen3-4B-Instruct-2507`
	`Qwen/Qwen3-8B`

DeepSeek	`deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`

Llama	`meta-llama/Llama-2-7b-hf`
	`meta-llama/Llama-3.1-8B`
	`meta-llama/Llama-3.1-8B-Instruct`
	`meta-llama/Llama-3.2-1B`
	`meta-llama/Llama-3.2-1B-Instruct`
	`meta-llama/Llama-3.2-3B-Instruct`

Mistral	`mistralai/Mistral-7B-v0.3` `mistralai/Mistral-7B-Instruct-v0.3`
Falcon	`tiiuae/falcon-7b-instruct`

GPT-2	`gpt2`
	`gpt2-medium`
	`gpt2-large`

TinyLlama	`TinyLlama/TinyLlama-1.1B-Chat-v1.0`

These values were chosen to create a deliberate asymmetry: the model favours persisting in runs of 1s ($\Delta\ell(11) = 3.0$, $\Delta\ell(111) = 2.5$) while making exits from 1s more likely than entries ($\Delta\ell(01) < 0$, $\Delta\ell(001) < 0$), and treats the two directions of alternation differently ($\Delta\ell(10) = 0$ vs. $\Delta\ell(01) = -1$). The stationary distribution at each temperature T is computed in closed form: for $k = 3$ via the Kimpton et al. (Kimpton et al., 2022) formula; for $k = 4$ by solving the linear stationarity system directly.

E.3 Model List

We initially considered the following 27 open-weight autoregressive language models (Hugging Face identifiers):

Binary sequence experiment. For all models, we computed the continuation probability $p_{\text{cont}}(L)$ as the average probability assigned to the correct next token of the pattern, averaged over cyclic shifts. We retained only models that achieved $p_{\text{cont}}(L) \geq 0.95$ for all context lengths $L \geq 20$ – i.e., models that demonstrably follow the instructed pattern – resulting in 17 models used in Figure 2; the remaining 10 models are excluded because a harmonic coordinate computed from a model that is not tracking the pattern would reflect instruction-following failure rather than a sequential bias, and the two are not distinguishable from the coordinate alone.

For each context length L and each cyclic shift of the pattern, we construct prompts for all q^{k-1} possible contexts $u \in \mathcal{A}^{k-1}$ (in the binary case, the four bigrams). Specifically, for a given shift, we take a prefix of length $L-2$ from the shifted pattern (wrapping cyclically) and then append the two tokens of the context u itself. The model is queried with this full prompt, and the softmax distribution over the next token is recorded. This procedure is repeated for all shifts, and the resulting probability vectors are averaged per context u. Because we explicitly construct prompts ending with every possible u, no context is left unobserved, even if the pattern never contains u naturally (e.g., the bigrams 00 and 11 in an alternating sequence). The resulting empirical transition matrix $\hat{T} \in \mathbb{R}^{q^{k-1} \times q}$ is then used to compute the stationary distribution via least squares, and the harmonic projections $\widehat{\gamma}_i(T)$ were computed with the eigenvectors from Example 4.7. Temperature was fixed at $T = 1.0$.

E.4 Relation between $\widehat{\Gamma}$ and the Pearson χ^2 statistic

At the aggregate level, $\widehat{\Gamma}$ is exactly related to the classical Pearson χ^2 statistic for departure from the uniform bigram distribution. Writing c_i for the nine empirical bigram counts, $N = \sum_i c_i$, $p_i = c_i/N$, and $u_i = 1/9$ for the uniform null,

$$X^2 \;=\; \sum_{i=1}^{9} \frac{(c_i - N/9)^2}{N/9} \;=\; 9N \sum_{i=1}^{9} \Big(p_i - \tfrac{1}{9}\Big)^2 \;=\; 9N\,\widehat{\Gamma}, \qquad \widehat{\Gamma} = \sum_i \widehat{\gamma}_i(1)^2 = \|p-u\|_2^2, \tag{14}$$

since the harmonic eigenvectors are an orthonormal basis for the mean-zero subspace and projection onto an orthonormal basis preserves squared Euclidean norm. So the scalar $\widehat{\Gamma}$ used for the win-rate correlation in Section 5.2 carries no more aggregate information than a chi-square departure from Nash. What the harmonic basis adds beyond this scalar is the *component-wise* spectrum $\{\widehat{\gamma}_i\}$: the individual coordinates separate current-move bias, previous-move bias, cyclic tendency, and repetition/anti-repetition into distinct, named, interpretable effects that a single aggregate statistic collapses. This is the same principle that motivates the finer-grained probing of (Bigelow et al., 2024): a pass/fail or uniform/non-uniform verdict is less useful than knowing *which* structure is present. In the RPS results of Section 5.2, the individual γ_i do not all move together across models, and it is this differential structure – not the aggregate $\widehat{\Gamma}$ alone – that lets the harmonic decomposition distinguish models that would look identical under a single non-uniformity score.

Rock-Paper-Scissors experiment. The 19 models that produced valid RPS play (each completing 500 games of 10 rounds) are listed below.

Model ID (Hugging Face)	Abbreviated name
`Qwen/Qwen3-8B`	Qwen3-8B
`Qwen/Qwen3-1.7B`	Qwen3-1.7B
`google/gemma-3-1b-pt`	Gemma-3-1B-pt
`google/gemma-3-4b-it`	Gemma-3-4B-it
`gpt2`	GPT-2
`gpt2-large`	GPT-2-Large
`meta-llama/Llama-2-7b-hf`	Llama-2-7B
`Qwen/Qwen2.5-7B`	Qwen2.5-7B
`Qwen/Qwen2.5-Coder-7B-Instruct`	Qwen2.5-Coder-7B-Inst
`meta-llama/Llama-3.2-1B-Instruct`	Llama-3.2-1B-Inst
`deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`	DeepSeek-R1-Qwen-14B
`gpt2-medium`	GPT-2-Medium
`google/gemma-2-9b-it`	Gemma-2-9B-it
`Qwen/Qwen2.5-Math-7B`	Qwen2.5-Math-7B
`meta-llama/Llama-3.1-8B-Instruct`	Llama-3.1-8B-Inst
`meta-llama/Llama-3.2-3B-Instruct`	Llama-3.2-3B-Inst
`TinyLlama/TinyLlama-1.1B-Chat-v1.0`	TinyLlama-1.1B-Chat
`Qwen/Qwen2.5-Math-7B-Instruct`	Qwen2.5-Math-7B-Inst
`Qwen/Qwen2.5-7B-Instruct`	Qwen2.5-7B-Inst

Each game consisted of 10 rounds. The model received a prompt containing the history of its own previous moves (as words, e.g., "rock, paper, ...") and was asked to output a single token R, P,

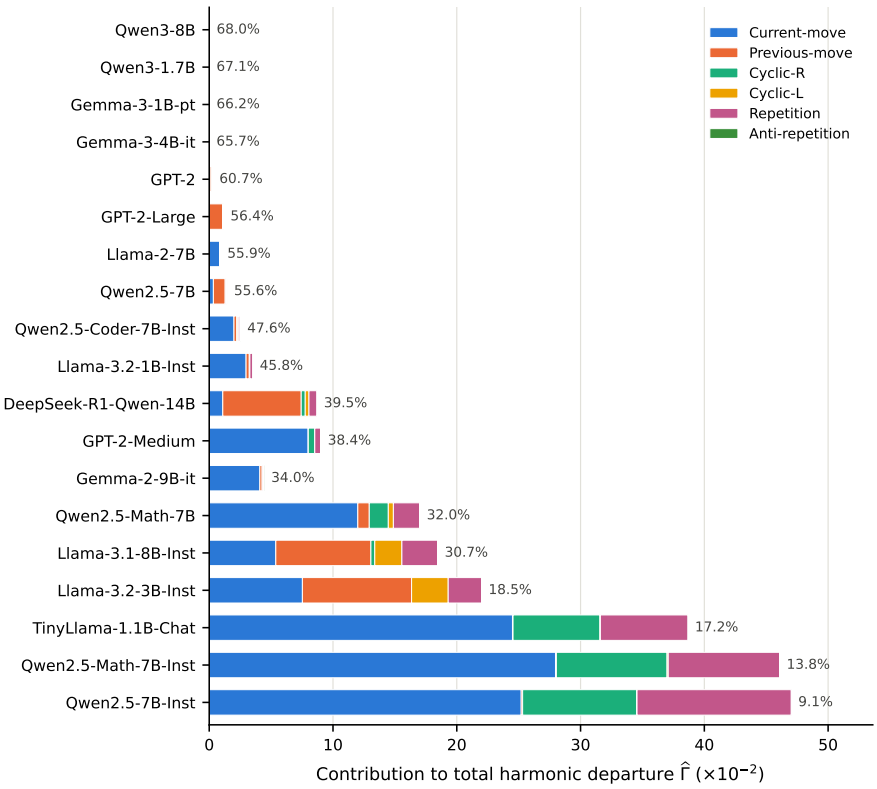

Figure 4: Graphical representation of the harmonic coordinates for all studied models with their respective win rates.

or S. The greedy adversary always played the move that beats the model's most probable move, with arbitrary tie-breaking when no single move dominated. The model's move probabilities were obtained by applying softmax to the raw logits at $T = 1.0$ and renormalising the probabilities of the tokens R, P, S to sum to one. For each model, we collected all 500×10 rounds and estimated a bigram transition matrix $\widehat{P} \in \mathbb{R}^{9\times9}$ over the nine bigram contexts of the previous move, computed as the empirical frequency of each transition. The stationary distribution $\widehat{\pi}$ was computed exactly (or via least squares if needed), and the harmonic spectrum was obtained by projecting $\widehat{\pi} - \pi_{\text{unif}}$ onto the eight eigenvectors derived from the ternary DFT basis.

Table 5 gives the full per-model coefficient table underlying Figure 4, sorted by win+tie rate.

Table 5: Harmonic spectral coefficients ($\times 10^{-2}$) sorted by Win+Tie% (i.e., non-loss percentage). Cells colored by absolute deviation from zero. A structural symmetry of the ternary de Bruijn graph forces $\gamma_1 = \gamma_3$ and $\gamma_2 = \gamma_4$ *exactly* for every model and so we show them as single columns.

Model	$\widehat{\gamma}_{1,3}$ (cur. move)	$\widehat{\gamma}_{2,4}$ (prev. move)	$\widehat{\gamma}_5$ (cyc-R)	$\widehat{\gamma}_6$ (cyc-L)	$\widehat{\gamma}_7$ (rep)	$\widehat{\gamma}_8$ (anti)	$\widehat{\Gamma}$ (total)	Win+Tie (%)
Qwen3-8B	-1.0	-0.5	+0.4	+0.1	+0.4	+0.0	< 0.1	68.0%
Qwen3-1.7B	-0.3	-0.8	-0.3	-0.2	+0.2	-0.3	< 0.1	67.1%
Gemma-3-1B-pt	+0.4	+0.2	+1.0	+0.5	-0.4	+0.4	< 0.1	66.2%
Gemma-3-4B-it	+0.7	+0.5	+0.3	+0.2	+0.1	-0.1	< 0.1	65.7%
GPT-2	+2.0	+2.8	-0.4	+0.4	+1.0	-0.4	+0.3	60.7%
GPT-2-Large	-0.9	+7.4	-0.5	+1.2	-1.5	+0.1	+1.2	56.4%
Llama-2-7B	+6.5	+0.2	+1.5	+0.7	+1.1	-1.1	+0.9	55.9%
Qwen2.5-7B	+4.1	-6.9	-2.3	-0.5	+3.0	+0.5	+1.4	55.6%
Qwen2.5-Coder-7B-Inst	+10.0	+3.5	+1.3	+3.0	-3.7	-1.3	+2.5	47.6%
Llama-3.2-1B-Inst	-12.2	+3.6	+1.3	-1.8	+4.8	+0.0	+3.5	45.8%
DeepSeek-R1-Qwen-14B	-7.4	+17.8	-5.7	-5.5	+8.0	-0.1	+8.7	39.5%
GPT-2-Medium	+20.0	-0.5	+7.2	-0.8	+6.9	-1.0	+9.0	38.4%
Gemma-2-9B-it	+14.3	-3.1	+1.0	-2.0	-2.3	+2.4	+4.4	34.0%
Qwen2.5-Math-7B	+24.5	+6.8	+12.4	+6.4	+14.6	-0.3	+17.0	32.0%
Llama-3.1-8B-Inst	-16.4	-19.6	-5.6	+14.8	+17.0	-0.3	+18.5	30.7%
Llama-3.2-3B-Inst	-19.4	+21.0	-0.4	-17.2	+16.4	+0.6	+22.0	18.5%
TinyLlama-1.1B-Chat	+35.0	-1.5	+26.5	-1.9	+26.6	-0.0	+38.6	17.2%
Qwen2.5-Math-7B-Inst	+37.4	+1.9	+29.9	+3.1	+30.0	+0.3	+46.0	13.8%
Qwen2.5-7B-Inst	+35.5	+2.4	+30.4	-0.3	+35.3	+0.1	+47.1	9.1%