Do Attention Maps in Neural Operators Learn the Green's Function?

Koffi Enakoutsa[1], Parker Bogard[1], and Prashant Jha[2]

[1]Department of Mathematics, University of California, Los Angeles
520 Portola Plaza, MS 6363, Box 951555
Los Angeles, CA 90095-1555, USA
koffi@math.ucla.edu

[2]Department of Mechanical Engineering
South Dakota School of Mines and Technology
Rapid City, SD 57701, USA

Abstract

Attention-based neural operators are increasingly used as fast surrogates for partial differential equations (PDEs), yet what their attention mechanism actually represents remains poorly understood. We observe that, for a *linear* PDE, the solution operator is an integral operator against the Green's function G, and that attention is itself a discretized kernel integral operator. On a grid, the discrete Green's function is exactly $\mathbf{A}^{-1}$, the inverse of the discretized differential operator, providing an *analytically known* ground truth—a rare luxury in interpretability research. We use this to ask a direct question: does a trained attention operator recover G, and when does it fail? We introduce a probing protocol that compares two read-outs of a trained model against $\mathbf{A}^{-1}$: the *raw attention weights* it displays, and the *effective linear operator* it computes (the map from forcing to solution the model actually realizes, independent of how the weights look). We score both with structure-aware alignment metrics. On controlled one-dimensional benchmarks (Poisson, screened Poisson, Helmholtz, a variable-coefficient operator, and non-self-adjoint advection–diffusion) we study the conditions under which alignment holds or breaks, including how recovery degrades as the target kernel becomes more oscillatory (large wavenumber, near resonance) or ill-conditioned. A central methodological finding is that operator fidelity does not imply attention interpretability: a model may solve the PDE accurately while its attention does not resemble G. In our small-scale experiments, the effective operator aligns with $\mathbf{A}^{-1}$ far more than the raw attention matrix does, and alignment degrades as the problem becomes harder, supporting this caution.

1 Introduction

Neural operators have emerged as effective data-driven surrogates for PDEs, learning a mapping between function spaces rather than the solution of a single instance [4, 3, 6]. Transformer-based operators, in particular, leverage attention to model long-range interactions across a discretized domain [10]. Despite strong empirical performance, these models are typically treated as black boxes, which limits trust in settings where physical correctness matters; a complementary line of work instead accepts the surrogate as a black box and corrects its residual error against the governing equations directly, whether in Bayesian inverse problems [7] or nonlinear variational boundary-value problems [8]. That line of work asks how to make a possibly-uninterpretable operator more accurate;

ours instead asks whether the operator's own internal attention mechanism is interpretable in the first place.

This paper asks a focused interpretability question. For a linear PDE $\mathcal{L}u = f$, the solution operator is the integral operator

$$u(x) = \int_\Omega G(x, y)\, f(y)\, \mathrm{d}y, \tag{1}$$

where G is the Green's function of $\mathcal{L}$ subject to the boundary conditions. In a single attention layer, each grid node x_i is a token, and the layer maps the token features (built from the sampled forcing $f(x_j)$ at every node, together with its coordinate x_j) to a weighted sum

$$\mathrm{Attn}(u)_i = \sum_j w_{ij}\, v_j, \tag{2}$$

a discretized kernel integral in which v_j is the value projection of the j-th token and the learned weights w_{ij} play the role of a kernel $\kappa(x_i, x_j)$. When the value projection is (or is close to) the identity, $v_j \approx f(x_j)$, and it is natural to read w_{ij} itself as an estimate $\widehat{G}(x_i, x_j)$ of the learned kernel, making the comparison with the true Green's function direct; Section 3 makes this precise as Eq. (5). It is therefore natural to conjecture that a trained attention operator should encode G. We test this conjecture directly.

The enabling observation is that the ground truth is exact and known. On a grid with discretized operator $\mathbf{A}$, the solution satisfies $\mathbf{Au} = \mathbf{f}$, hence $\mathbf{u} = \mathbf{A}^{-1}\mathbf{f}$; the matrix $\mathbf{A}^{-1}$ is the *discrete Green's function*, with $(\mathbf{A}^{-1})_{ij} \approx h\, G(x_i, x_j)$ under a standard quadrature convention—we prove this identity exactly for the benchmarks we study in Section 4 (Theorem 4.1). Unlike attention interpretability in natural language, where the "correct" attention map is undefined, here we know precisely what a faithful attention map should look like. This makes linear PDEs an unusually clean testbed for studying what attention learns.

Problem. The problem we address is whether the attention mechanism of a neural operator, trained to solve a linear PDE, actually *represents* the operator it has learned to apply—concretely, whether its attention map encodes the Green's function G. A subtlety makes this non-trivial: a model can be an accurate surrogate (small solution error) for two different reasons—either because its attention has discovered the kernel, or because some other internal pathway computes the solution while the attention does something unrelated. Solution accuracy alone cannot tell these apart, so the interpretability question is *not* settled by performance and must be tested directly against the true kernel. A further subtlety concerns fairness of the comparison itself: when the value projection is not the identity, $v_j \neq f(x_j)$, and the raw weights w_{ij} are not directly comparable to $G(x_i, x_j)$, since a rescaling of v_j can be compensated by a rescaling of w_{ij} without changing the model's output. Section 5 addresses this by scoring w_{ij} against G only up to a single global affine transform (Eq. (11)), the minimal correction that any linear reparameterization of the value pathway can induce; this keeps the comparison fair while still testing whether the *structure* of G is visible in attention.

Approach. Our approach exploits the exactness of the discrete ground truth. We (i) fix controlled one-dimensional linear benchmarks whose Green's function—equivalently the discrete inverse $\mathbf{A}^{-1}$—is known analytically; (ii) extract two distinct objects out of a trained attention operator, the raw attention matrix $\mathbf{W}$ (what the model *displays*) and its effective linear operator $\mathbf{M}$ (what the model *computes*); and (iii) score each against $\mathbf{A}^{-1}$ with structure-aware metrics that separate symmetry,

locality, singularity, and oscillation. Comparing the two readouts converts a qualitative debate into a measurable test and, as the theory of Section 4 shows, forces a clean separation between operator fidelity and attention interpretability.

Contributions.

1. A **probing protocol** that extracts, from a trained attention operator, both (a) the raw attention weight matrix and (b) the model's effective linear operator, and compares each against $\mathbf{A}^{-1}$ (Sec. 5).

2. A set of **structure-aware alignment metrics**—relative L^2 error, symmetry error, off-diagonal decay, peak/singularity localization, and structural correlation—that quantify recovery of G beyond a single scalar.

3. A **controlled empirical study** establishing when alignment holds and when it breaks, including the methodological caution that operator fidelity does not entail attention interpretability (Sec. 7).

This is explicitly *not* a new architecture. We introduce no architectural novelty and claim no accuracy improvement, the contribution is diagnostic and interpretive.

Hypotheses. The theory and experiments are organized around four hypotheses, which we evaluate as follows:

H1. *(Operator fidelity is achievable.)* With sufficient depth and data, the model's effective linear operator $\mathbf{M}$ aligns with the discrete Green's function $\mathbf{A}^{-1}$, and this alignment improves with capacity and sample size.

H2. *(Attention recovers the kernel.)* The raw attention matrix $\mathbf{W}$, suitably aligned, recovers the structure of G—the naive conjecture suggested by the kernel-integral reading of attention.

H3. *(Difficulty degrades recovery.)* Alignment with $\mathbf{A}^{-1}$ degrades as the problem becomes harder—oscillatory (large wavenumber, near resonance), strongly localized, or strongly advective.

H4. *(Fidelity does not imply interpretability.)* Operator fidelity ($\mathbf{M} \approx \mathbf{A}^{-1}$) does not entail attention interpretability ($W \approx G$): a model may compute the correct operator while its attention map does not resemble the kernel.

Our central results are negative for H2 and positive for H1 and H4: the attention probe *provably* cannot reach $\mathbf{A}^{-1}$ under the alignment the architecture permits (Proposition 4.8), whereas the operator probe can (Theorem 4.10), so the two are formally separated (Corollary 4.11). H1 and H3 are then characterized empirically (Sec. 7).

Organization. Section 2 situates the question among neighboring lines of work. Section 3 fixes the continuous problem, its discretization, and the attention layer we probe. Section 4 develops the theory: the exact discrete Green's function, the separation between operator fidelity and attention interpretability, and the identifiability of the operator probe. Section 5 specifies the probing protocol and alignment metrics; Section 6 describes the five benchmarks, data generation, and training; Section 7 reports the empirical findings; Section 8 draws out the methodological implications and scope; and Section 9 concludes.

2 Related Work

Before specifying our protocol, we situate it among four neighboring lines of work—operator learning, transformer-based operators, methods that learn Green's functions directly, and the broader debate on whether attention is explanatory—and note that the question we raise, what the attention *inside* an operator encodes, falls in the gap between them.

Operator learning. Operator learning approximates a solution operator $\mathcal{G} : a \mapsto u$ between function spaces, so a new instance is solved by a forward pass rather than a numerical solve. Representative approaches include DeepONet [2], the Fourier Neural Operator [3], and the general neural-operator framework of [4]; by contrast, physics-informed networks [5] fit a single instance by penalizing the PDE residual. These methods motivate our object of study but do not ask what their internal representations encode.

Transformer-based operators. Attention has been adapted to operator learning by treating mesh nodes as tokens. The Galerkin Transformer [9] is closest to our concerns: it proves that the softmax normalization is sufficient but not necessary, and that a linear attention variant is layerwise comparable to a Petrov–Galerkin projection—an analysis of the value pathway that complements our separation result for the raw softmax map. Later architectures scale attention to large meshes and complex geometries, e.g. OFormer [10], GNOT [11], and Transolver [12], which applies attention to learnable physical "slices" in linear complexity. These are all *different* architectures for building operators; we instead *probe* what the attention inside such a model represents.

Learning Green's functions. A distinct line learns the Green's function explicitly. [13] train rational networks to recover G from forcing–response pairs and read off symmetries, conservation laws, and singularities; DeepGreen [14] targets nonlinear boundary value problems. Notably, [13] report that operator-learning methods can predict solutions accurately yet fail to recover G itself—an empirical precursor to our formal separation of operator fidelity from kernel recovery (Section 4). The contrast with our work is one of aim: these methods design architectures to learn G, whereas we probe whether the attention of a standard operator encodes it.

Attention as a kernel. [15] recast attention [1] as a kernel smoother—a weighted average with similarity-based weights—which is exactly the Nadaraya–Watson view we formalize in Proposition 4.5. This viewpoint makes the softmax normalization, and hence the row-stochastic constraint behind Proposition 4.8, intrinsic rather than incidental.

Interpreting attention. Whether attention weights "explain" a model has been debated in natural language processing [16, 17], complicated by the absence of any ground-truth attention map. Our setting removes that ambiguity: for a linear PDE the faithful kernel is analytically known ($\mathbf{A}^{-1}$, Theorem 4.1), turning a qualitative debate into a quantitative test against an exact target.

Green's functions and the discrete target. G encodes interpretable structure—symmetry for self-adjoint $\mathcal{L}$, off-diagonal decay for coercive/local operators, oscillation for indefinite ones such as Helmholtz—which our metrics target. Discretization replaces G by the matrix inverse $\mathbf{A}^{-1}$; the formal setup, and the exact relationship between $\mathbf{A}^{-1}$ and G, are developed in Section 3 and Section 4.

3 Setup and Notation

To turn the informal question of the previous sections into a measurable one, we now fix the continuous problem, its discretization, and the attention layer we probe; this is the machinery on which both the theory and the experiments rest. We work on $\Omega = (0,1)$ with homogeneous Dirichlet conditions. Let $\mathcal{L}$ be a linear second-order operator and consider

$$\mathcal{L}u = f \ \text{ in } \Omega, \qquad u(0) = u(1) = 0, \tag{3}$$

with Green's function G satisfying $\mathcal{L}_x G(\cdot, y) = \delta(\cdot - y)$ and the boundary conditions, so that $u(x) = \int_0^1 G(x,y) f(y)\,\mathrm{d}y$.

Discretization. Let $h = 1/(N+1)$ and $x_i = ih$, $i = 0, 1, \ldots, N, N+1$, so that $x_0 = 0$ and $x_{N+1} = 1$ are the two boundary nodes at which the homogeneous Dirichlet conditions $u_0 = 0$ and $u_{N+1} = 0$ are imposed; the N interior nodes $x_1, \ldots, x_N$ carry the unknowns. A second-order finite-difference discretization on the interior nodes yields a matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ and a linear system $\mathbf{Au} = \mathbf{f}$ for nodal vectors $\mathbf{u}, \mathbf{f} \in \mathbb{R}^N$ (with $u_i \approx u(x_i)$, $f_i = f(x_i)$). For the Dirichlet Laplacian $\mathcal{L}u = -u''$,

$$\mathbf{A} = \tfrac{1}{h^2}\,\mathrm{tridiag}(-1, 2, -1). \tag{4}$$

We call $\mathbf{A}^{-1}$ the *discrete Green's function* and write $G^h_{ij} = (\mathbf{A}^{-1})_{ij}$.

Attention operator. Grid nodes are tokens carrying the pair (f_i, x_i): the sampled forcing value and the spatial coordinate, the latter entering through a coordinate positional encoding. A single-head attention layer maps the token features through learned query, key, and value projections and produces, for attention weights $\mathbf{W} \in \mathbb{R}^{N\times N}$,

$$\mathrm{Attn}(\mathbf{u})_i = \sum_{j=1}^{N} W_{ij} (\mathbf{Vu})_j, \qquad W_{ij} = \frac{\exp(s_{ij})}{\sum_\ell \exp(s_{i\ell})}, \tag{5}$$

where $\mathbf{V} \in \mathbb{R}^{N\times N}$ is the value projection applied to the token vector $\mathbf{u}$ (so $(\mathbf{Vu})_j$ plays the role of v_j in Eq. (2)), s_{ij} are the scaled query–key scores, and the normalization $W_{ij} = \exp(s_{ij})/\sum_\ell \exp(s_{i\ell})$ is the standard *softmax* attention weighting: it makes each row of $\mathbf{W}$ a probability vector, i.e. $\mathbf{W}$ is row-stochastic (Lemma 4.7), a property used throughout Section 4. The layer's output is $\mathbf{O}\,\mathrm{Attn}(\mathbf{u})$ for an output projection $\mathbf{O} \in \mathbb{R}^{N\times N}$, added to the token vector $\mathbf{u}$ through a residual connection, so the full layer computes $\mathbf{u} + \mathbf{O}\,\mathrm{Attn}(\mathbf{u})$. We analyze this layer under the following design assumption.

Assumption 3.1 (Coordinate attention)**.** The scores s_{ij} depend only on the coordinates (x_i, x_j), not on the forcing $\mathbf{f}$; equivalently, the attention weights are fixed once training is complete and do not vary with the input forcing at inference time (the input-dependent case, in which $\mathbf{W}$ varies with $\mathbf{f}$, is treated in Remark 4.12). Under this assumption $\mathbf{W}$ is a fixed matrix, the layer $\mathbf{u} \mapsto \mathbf{u} + \mathbf{O}\,\mathrm{Attn}(\mathbf{u})$ acts *linearly* on $\mathbf{f}$, and the layer's output can be written as the action of a single *effective linear operator*

$$\mathbf{u} + \mathbf{O}\,\mathrm{Attn}(\mathbf{u}) = \mathbf{Mf}, \qquad \mathbf{M} = \mathbf{I} + \mathbf{O}\,\mathbf{W}\,\mathbf{V} \in \mathbb{R}^{N\times N}, \tag{6}$$

where the identity $\mathbf{I}$ comes from the residual connection: it is kept explicit here (rather than folded into $\mathbf{O}, \mathbf{V}$) because it reappears as a visible artifact in the recovered kernels of Section 7 and is used explicitly in the proof of Theorem 4.10.

Remark 3.2. Assumption 3.1 is the interpretable, content-independent regime in which reading the attention map is meaningful; it holds for our primary coordinate-attention model, while the general input-dependent case is treated in Remark 4.12.

We compare two probes against $\mathbf{A}^{-1}$: the *attention probe* reads $\mathbf{W}$, allowing for a *global affine map*—a single scalar rescaling and shift applied to every entry of $\mathbf{W}$, $\mathbf{W} \mapsto a\mathbf{W} + b\mathbf{1}\mathbf{1}^\top$ for scalars a, b (made precise in Eq. (11) and used in Theorem 4.10)—and the *operator probe* reads $\mathbf{M}$ directly, with no such allowance needed.

4 Theory

With the setup of Section 3 in place, we establish the three facts on which the empirical study rests: that $\mathbf{A}^{-1}$ is an exact and interpretable ground truth; that softmax attention is structurally barred from matching it while the effective operator is not (our central separation, H2 versus H4); and that the operator probe is identifiable, with an error controlled by the model's solution error. The subsections proceed in that order.

4.1 The discrete Green's function

Theorem 4.1 (Exact discrete Green's function for the Dirichlet Laplacian)**.** *Let* $\mathbf{A}$ *be given by* (4) *with* $h = 1/(N+1)$*. Then for all* $1 \le i, j \le N$*,*

$$(\mathbf{A}^{-1})_{ij} = h\, G(x_i, x_j), \qquad G(x,y) = \begin{cases} x(1-y), & x \le y, \\ y(1-x), & x \ge y, \end{cases} \tag{7}$$

where G *is the Green's function of* $-u''$ *with Dirichlet conditions. The identity is exact.*

Proof. Fix j and let $\mathbf{g} := \mathbf{A}^{-1}\mathbf{e}_j$, so $\mathbf{A}\mathbf{g} = \mathbf{e}_j$, i.e.

$$-g_{i-1} + 2g_i - g_{i+1} = h^2\delta_{ij}, \qquad g_0 = g_{N+1} = 0. \tag{8}$$

For $i \neq j$ the right-hand side vanishes, so this is the homogeneous second-difference recurrence $g_{i+1} - 2g_i + g_{i-1} = 0$, whose general solution is a *linear function of the grid index* i, $g_i = c_1 + c_2 i$; with the boundary conditions $\mathbf{g}$ is therefore piecewise linear in i on each side of the impulse location j,

$$g_i = \alpha\, i \ (i \le j), \qquad g_i = \beta(N+1-i) \ (i \ge j). \tag{9}$$

Continuity at $i = j$ gives $\alpha j = \beta(N+1-j)$. Evaluating (8) at $i = j$, the second difference of a piecewise linear function equals the slope jump: $\alpha - (-\beta) = \alpha + \beta = h^2$. Solving, $\alpha = h^2(N+1-j)/(N+1)$, so for $i \le j$, $(\mathbf{A}^{-1})_{ij} = \alpha i = h^2 i(N+1-j)/(N+1)$. Since $x_i \le x_j$ gives $h\,G(x_i, x_j) = h\,x_i(1-x_j) = i(N+1-j)/(N+1)^3 = h^2 i(N+1-j)/(N+1)$, the two agree; the case $i \ge j$ is symmetric. □

Remark 4.2. Theorem 4.1 is the analytic basis for using $\mathbf{A}^{-1}$ as ground truth: the discrete and continuous Green's functions agree at the nodes up to the factor h with zero error, explaining the $\sim 10^{-15}$ discrepancy observed numerically.

Proposition 4.3 (Stability and consistency, coercive self-adjoint case)**.** *For* $\mathcal{L}u = -(a(x)u')' + c(x)u$ *with* $a, c \in C^2$*,* $a \ge a_0 > 0$*,* $c \ge 0$*, discretized by the* conservative scheme*—the central-difference, flux-form discretization* $-\big(a_{i+1/2}(u_{i+1} - u_i) - a_{i-1/2}(u_i - u_{i-1})\big)/h^2 + c_i u_i$*, with* $a_{i\pm1/2} = a(x_i \pm h/2)$*, that discretely mimics conservation of the flux* $a(x)u'(x)$*—the matrix* $\mathbf{A}$ *is symmetric positive definite with* $\|\mathbf{A}^{-1}\|_2 \le C$ *uniformly in* N*, and for each interior* $y = x_j$ *the rescaled column* $h^{-1}(\mathbf{A}^{-1})_{\cdot j}$ *converges to* $G(\cdot, y)$ *as* $N \to \infty$*.*

Proof sketch. Symmetry follows from conservative differencing; coercivity ($a \geq a_0$, $c \geq 0$) gives positive definiteness and the uniform bound $\|\mathbf{A}^{-1}\|_2 \leq 1/(8a_0)$ (Appendix A, Lemma A.2). The uniform convergence of the rescaled discrete Green's function to $G(\cdot, x_j)$, including *through* the diagonal where G has a slope discontinuity, is established in Appendix A (Theorem A.5) via the equivalence of the conservative scheme with P_1 finite elements, at rate $O(h)$ in the discrete maximum norm. □

Remark 4.4. Proposition 4.3 covers Poisson, screened Poisson, and the variable-coefficient operator. Helmholtz is symmetric but indefinite (invertible away from the resonances of Proposition 4.19); advection–diffusion is non-symmetric (Proposition 4.17).

4.2 Attention as a self-normalized kernel smoother

Proposition 4.5. *Under Assumption 3.1, if $s_{ij} = \log \kappa(x_i, x_j)$ with $\kappa > 0$, the weights are Nadaraya–Watson weights $W_{ij} = \kappa(x_i, x_j)/\sum_\ell \kappa(x_i, x_\ell)$, and for smooth κ and smooth $v(\cdot)$ with nodal values $v_j = v(x_j)$,*

$$(\mathbf{W}\mathbf{v})_i = \frac{\sum_j \kappa(x_i, x_j) v_j}{\sum_\ell \kappa(x_i, x_\ell)} = \frac{\int_0^1 \kappa(x_i, y) v(y)\, \mathrm{d}y}{\int_0^1 \kappa(x_i, y)\, \mathrm{d}y} + O(h^2). \tag{10}$$

In particular $\mathbf{W}$ *preserves constants:* $\mathbf{W}\mathbf{1} = \mathbf{1}$.

Proof. The first identity is softmax with $s_{ij} = \log \kappa$. The midpoint rule gives $O(h^2)$ quadratures of numerator and denominator; the ratio inherits the rate since the denominator is bounded below. Constant preservation is row-stochasticity (Lemma 4.7). □

Remark 4.6. This is the continuous origin of Proposition 4.8: softmax attention is an averaging operator and maps constants to constants, whereas $\int G(x, y)\mathrm{d}y$ solves $\mathcal{L}u = 1$ and is non-constant.

4.3 Operator fidelity versus attention interpretability

Let $\mathcal{W} = \{\mathbf{W} : W_{ij} \geq 0,\ \mathbf{W}\mathbf{1} = \mathbf{1}\}$ be the row-stochastic matrices and

$$\mathcal{A} = \{a\mathbf{W} + b\,\mathbf{1}\mathbf{1}^\top : a, b \in \mathbb{R},\ \mathbf{W} \in \mathcal{W}\} \tag{11}$$

the set an attention probe can match under a global affine alignment.

Lemma 4.7. *Any* $\mathbf{W}$ *from* (5) *satisfies* $\mathbf{W} \in \mathcal{W}$.

Proof. $W_{ij} = e^{s_{ij}}/\sum_\ell e^{s_{i\ell}} > 0$ and $\sum_j W_{ij} = 1$. □

Proposition 4.8 (The attention probe cannot reach $\mathbf{A}^{-1}$)**.** *If the row sums of $\mathbf{A}^{-1}$ are not all equal, then $\mathbf{A}^{-1} \notin \mathcal{A}$; no softmax attention matrix matches $\mathbf{A}^{-1}$ even after global affine alignment. This holds for the Dirichlet Laplacian, and whenever the discrete constant-load solution is non-constant.*

Proof. If $\mathbf{A}^{-1} = a\mathbf{W} + b\mathbf{1}\mathbf{1}^\top$, then $\mathbf{A}^{-1}\mathbf{1} = (a + bN)\mathbf{1}$, so all rows of $\mathbf{A}^{-1}$ sum equally, i.e. $\mathbf{A}^{-1}\mathbf{1}$ (the discrete solution of $\mathcal{L}u = 1$) is constant. For $-u'' = 1$ this is the sampling of $\frac{1}{2}x(1-x)$, non-constant; contradiction. □

Remark 4.9 (Why a per-row affine alignment is excluded). The obstruction in Proposition 4.8 is specific to a *single* global pair (a, b). A per-row affine alignment $\hat{W}_{i\cdot} = a_i W_{i\cdot} + b_i \mathbf{1}^\top$ introduces N independent shifts b_i, which can match the row sums of $\mathbf{A}^{-1}$ one row at a time: for any non-constant target row r_i one may take $b_i \le \min_j r_{ij}$ and $a_i = \sum_j r_{ij} - N b_i > 0$, recovering r_i exactly from a probability vector. A row-wise alignment therefore voids the constant-preservation constraint and makes the comparison vacuous. We consequently fix the global family $\mathcal{A}$ of (11) as the only alignment for the attention probe (Sec. 5); it is also the family the architecture actually realizes, since softmax normalization is identical across rows and the value/output projections supply a single overall scale.

Theorem 4.10 (Representational adequacy of the effective operator)**.** *Let $\mathbf{W} \in \mathcal{W}$ be invertible. Then there exist linear projections $\mathbf{V}, \mathbf{O}$ with $\mathbf{M} = \mathbf{OWV} = \mathbf{A}^{-1}$ exactly (no residual); the operator probe faces no obstruction of the type in Proposition 4.8. The same holds for the residual-inclusive form $\mathbf{M} = \mathbf{I} + \mathbf{OWV}$ of Eq.* (6) *actually used by the architecture.*

Proof. Without a residual, take $\mathbf{V} = \mathbf{I}$, $\mathbf{O} = \mathbf{A}^{-1}\mathbf{W}^{-1}$; then $\mathbf{M} = \mathbf{A}^{-1}\mathbf{W}^{-1}\mathbf{W} = \mathbf{A}^{-1}$. With the residual of Eq. (6), instead take $\mathbf{O} = (\mathbf{A}^{-1} - \mathbf{I})\mathbf{W}^{-1}$, so that $\mathbf{M} = \mathbf{I} + \mathbf{OWV} = \mathbf{I} + (\mathbf{A}^{-1} - \mathbf{I}) = \mathbf{A}^{-1}$. □

Corollary 4.11 (Formal separation, H4)**.** *Under Assumption 3.1, for operators satisfying Proposition 4.8 (all five benchmarks) there are parameters with $\mathbf{M} = \mathbf{A}^{-1}$ exactly while the attention probe is bounded away from $\mathbf{A}^{-1}$, since $\mathbf{A}^{-1} \notin \mathcal{A}$. A faithful operator probe with a poorly aligned attention probe is therefore the generic, expected outcome, not a training failure.*

Remark 4.12. Without Assumption 3.1, $\mathbf{W}(\mathbf{f})$ depends on $\mathbf{f}$; the attention probe reports $\overline{\mathbf{W}} = \mathbb{E}_{\mathbf{f}}[\mathbf{W}(\mathbf{f})]$, a convex combination of row-stochastic matrices, hence still in $\mathcal{W}$, so Proposition 4.8 applies verbatim. The operator probe reports the data-averaged Jacobian, quantified next.

4.4 Identifiability of the operator probe

Let $g : \mathbb{R}^N \to \mathbb{R}^N$ be the trained model, mapping a forcing vector to a predicted solution vector, and let $\mathbf{f}$ have zero mean and covariance $\mathbf{\Sigma} = \mathbb{E}[\mathbf{f}\mathbf{f}^\top] \succ 0$. The least-squares operator probe is $\mathbf{M} = \arg\min_{\mathbf{M}'} \mathbb{E}\|g(\mathbf{f}) - \mathbf{M}'\mathbf{f}\|_2^2$, the population version of the regression used in the experiments; write $\varepsilon^2 = \mathbb{E}\|g(\mathbf{f}) - \mathbf{A}^{-1}\mathbf{f}\|_2^2$.

Theorem 4.13 (Identifiability and error bound)**.** *If $\mathbf{\Sigma} \succ 0$, the minimizer is unique, $\mathbf{M} = \mathbb{E}[g(\mathbf{f})\mathbf{f}^\top]\mathbf{\Sigma}^{-1}$, and*

$$\|\mathbf{M} - \mathbf{A}^{-1}\|_F \;\le\; \frac{\varepsilon\sqrt{\operatorname{tr}\mathbf{\Sigma}}}{\sigma_{\min}(\mathbf{\Sigma})}. \tag{12}$$

Proof. Strict convexity ($\mathbf{\Sigma} \succ 0$) gives the unique stationary point $\mathbf{M} = \mathbb{E}[g(\mathbf{f})\mathbf{f}^\top]\mathbf{\Sigma}^{-1}$. With $\mathbf{r}(\mathbf{f}) = g(\mathbf{f}) - \mathbf{A}^{-1}\mathbf{f}$ ($\mathbb{E}\|\mathbf{r}\|^2 = \varepsilon^2$), $\mathbf{M} = \mathbf{A}^{-1} + \mathbb{E}[\mathbf{r}\mathbf{f}^\top]\mathbf{\Sigma}^{-1}$, so $\|\mathbf{M} - \mathbf{A}^{-1}\|_F \le \|\mathbb{E}[\mathbf{r}\mathbf{f}^\top]\|_F \|\mathbf{\Sigma}^{-1}\|_2$. By convexity of the norm and Cauchy–Schwarz, $\|\mathbb{E}[\mathbf{r}\mathbf{f}^\top]\|_F \le \mathbb{E}[\|\mathbf{r}\|\|\mathbf{f}\|] \le \varepsilon\sqrt{\operatorname{tr}\mathbf{\Sigma}}$, and $\|\mathbf{\Sigma}^{-1}\|_2 = 1/\sigma_{\min}(\mathbf{\Sigma})$. □

Corollary 4.14. *If $g(\mathbf{f}) = \mathbf{B}\mathbf{f}$ for a matrix $\mathbf{B}$, then $\mathbf{M} = \mathbf{B}$ for every $\mathbf{\Sigma} \succ 0$; if also $\varepsilon = 0$ then $\mathbf{M} = \mathbf{A}^{-1}$. The operator probe is exact for a linear model that solves the PDE.*

Remark 4.15. Equation (12) both justifies the probe and bounds its reach: with modal energies decaying as m^{-2s} (our $s = 1.2$), $\sigma_{\min}(\mathbf{\Sigma})$ is the smallest mode energy, so the operator is identified on the well-excited low modes (where G concentrates) and weakly on high modes. With Proposition 4.8, Theorem 4.13 completes the account of H4: a small ε forces $\mathbf{M} \approx \mathbf{A}^{-1}$, while no parameters force $\mathbf{W} \approx \mathbf{A}^{-1}$.

4.5 Symmetry and the Péclet number

Proposition 4.16. *If the scheme is symmetric ($\mathbf{A} = \mathbf{A}^\top$), then $\mathbf{A}^{-1} = (\mathbf{A}^{-1})^\top$.*

Proof. $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1} = \mathbf{A}^{-1}$. □

Proposition 4.17 (Advection breaks symmetry linearly in the Péclet number)**.** *For $-\nu u'' + b\,u'$ with central differences, $\mathbf{A} = \nu\mathbf{L} + b\mathbf{S}$ with $\mathbf{L} = \frac{1}{h^2}\mathrm{tridiag}(-1, 2, -1) \succ 0$ symmetric and $\mathbf{S} = \frac{1}{2h}\mathrm{tridiag}(-1, 0, 1)$ skew, $\mathbf{S}^\top = -\mathbf{S}$. Then $\mathbf{A}^{-1}$ is symmetric iff $b = 0$, and for $p = b/\nu$, $\mathbf{A}^{-1} = \frac{1}{\nu}\mathbf{L}^{-1} - \frac{p}{\nu}\mathbf{L}^{-1}\mathbf{S}\mathbf{L}^{-1} + O(p^2)$, whose second term is antisymmetric; the relative asymmetry grows linearly in p.*

Proof. Skewness of $\mathbf{S}$ is entrywise. Factor $\mathbf{A} = \nu\mathbf{L}(\mathbf{I} + p\mathbf{L}^{-1}\mathbf{S})$ and expand: $\mathbf{A}^{-1} = (\mathbf{I} - p\mathbf{L}^{-1}\mathbf{S} + O(p^2))\frac{1}{\nu}\mathbf{L}^{-1}$. The leading term is symmetric; $(\mathbf{L}^{-1}\mathbf{S}\mathbf{L}^{-1})^\top = \mathbf{L}^{-1}\mathbf{S}^\top\mathbf{L}^{-1} = -\mathbf{L}^{-1}\mathbf{S}\mathbf{L}^{-1}$ is antisymmetric and is the first-order asymmetric part, of size $O(p)$ relative to the symmetric part. $\mathbf{A}^{-1}$ is symmetric iff this vanishes, i.e. $b = 0$. □

Remark 4.18. This makes the symmetry-error metric a genuine discriminator: near zero for the four self-adjoint benchmarks, growing with p for advection–diffusion. A faithful recovery must reproduce the asymmetry, as the operator probe does.

4.6 Conditioning and the difficulty trend

The eigenvalues of $\mathbf{L}$ are $\lambda_m = \frac{4}{h^2}\sin^2(\frac{m\pi h}{2})$.

Proposition 4.19 (Helmholtz: near-resonant ill-conditioning)**.** *For $-u'' - k^2u$, $\mathbf{A} = \mathbf{L} - k^2\mathbf{I}$ has $\|\mathbf{A}^{-1}\|_2 = 1/\operatorname{dist}(k^2, \{\lambda_m\})$, which diverges as k^2 approaches an eigenvalue; with Theorem 4.13 the error bound is amplified near resonance and $\mathbf{A}^{-1}$ becomes strongly oscillatory.*

Proof. $\mathbf{A}$ shares the eigenvectors of $\mathbf{L}$ with eigenvalues $\lambda_m - k^2$, so $\|\mathbf{A}^{-1}\|_2 = 1/\min_m|\lambda_m - k^2|$; near $k^2 = \lambda_{m^\star}$ the column is dominated by $\boldsymbol{\phi}_{m^\star}/(\lambda_{m^\star} - k^2)$. □

Proposition 4.20 (Screened Poisson: localization, not ill-conditioning)**.** *For $-u'' + \kappa^2u$, $\mathbf{A} = \mathbf{L} + \kappa^2\mathbf{I}$ has $\sigma_{\min} = \lambda_1 + \kappa^2$ (bounded conditioning as κ grows), while $G_\kappa(x, y) \sim \frac{1}{2\kappa}e^{-\kappa|x-y|}$ has support length $O(1/\kappa)$.*

Proof. Immediate from the eigenvalues and the screened Green's function on the line, up to boundary corrections exponentially small in κ. □

Remark 4.21. Two distinct difficulty mechanisms, matching the experiments: Helmholtz is hard through conditioning (alignment collapses near resonance, large k); screened Poisson is well-conditioned but its kernel concentrates on a vanishing band, so the structural-correlation metric loses off-diagonal signal-to-noise—hence its noisy, non-monotone trend.

5 Method

The theory of Section 4 tells us *what* to measure—the gap between what attention displays and what the operator computes. This section specifies *how* we measure it: the model we probe, the two read-outs, and the structure-aware metrics that score them against $\mathbf{A}^{-1}$.

An intentionally interpretable operator. We treat each grid point as a token carrying its sampled forcing value f_i and its coordinate x_i; the coordinate enters through a positional encoding (a small fixed set of sinusoidal or polynomial features of x_i, concatenated to f_i) so the query/key projections can depend on position. Our primary model is a *single-head* attention operator, Eq. (5), with one or two layers: query, key, and value projections are each a single linear map $\mathbb{R}^{d_{\text{model}}} \to \mathbb{R}^{d_{\text{model}}}$ with $d_{\text{model}} = 64$ (Sec. 6), so $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{N \times d_{\text{model}}}$ and the score matrix is $\mathbf{S} = \mathbf{Q}\mathbf{K}^\top / \sqrt{d_{\text{model}}}$; the value and output projections $\mathbf{V}, \mathbf{O}$ are likewise single linear maps (no hidden nonlinearity), deliberately kept simple so the attention matrix is as directly readable as possible, with a residual connection around the layer (Eq. (6)). We also train an *expressive* variant (multi-head, wider feed-forward block) to test whether added capacity hides the kernel structure rather than exposing it; Table 2 reports the effect of depth and number of heads.

Probe 1: attention readout. We extract the attention weight matrix $\mathbf{W}$ (averaged over a held-out set of forcings, and also examined per-forcing). Because softmax rows are normalized and the value/output projections supply a single overall scale, we compare $\mathbf{W}$ to $\mathbf{A}^{-1}$ *up to one global affine transform* $\mathbf{W} \mapsto a\mathbf{W} + b\,\mathbf{1}\mathbf{1}^\top$ $(a, b \in \mathbb{R})$, fit by least squares and reported, rather than expecting raw equality. The intuition is that this is precisely the freedom the architecture itself has left after normalization: softmax already fixes each row to sum to one, so the only remaining degrees of freedom by which a linear rescaling of the value/output projections can alter the visible weights are one global scale a and one global shift b applied uniformly to every entry—not an independent shift per row. This is exactly the family $\mathcal{A}$ of (11), so the comparison is the one for which Proposition 4.8 certifies an obstruction. We deliberately do *not* use a per-row affine map: as shown in Remark 4.9, N independent row shifts can match the row sums of $\mathbf{A}^{-1}$ one row at a time, voiding the constant-preservation constraint that drives Proposition 4.8 and rendering the comparison vacuous. This probe asks: is the kernel structure of G *visible* in attention under the single alignment the architecture actually permits?

Probe 2: effective-operator recovery. Independently of what the weights display, we recover what the model *computes*. For the linear-data regime, we estimate the model's effective linear operator $\mathbf{M}$ column by column from its impulse responses: feeding each basis forcing $\mathbf{e}_j$ $(j = 1, \ldots, N)$ through the trained model and reading off the output as the j-th column of $\mathbf{M}$; this is equivalent to the Jacobian $\partial\mathbf{u}/\partial\mathbf{f}$ for a linear (or linearized) model and coincides with the least-squares regression probe of Theorem 4.13 in the population limit. We use the direct impulse-response construction in the reported experiments, since the trained models here are linear in $\mathbf{f}$ under Assumption 3.1, and assemble $\mathbf{M} \approx \mathbf{A}^{-1}$. This probe is robust to reparameterization and measures functional fidelity rather than visual resemblance.

A mismatch is a finding. The gap between Probe 1 and Probe 2 is itself informative: a model can have $\mathbf{M} \approx \mathbf{A}^{-1}$ (accurate operator) while $\mathbf{W}$ does not resemble $\mathbf{A}^{-1}$ (uninterpretable attention). We report this gap explicitly rather than treating it as failure.

Alignment metrics. Given an estimated kernel matrix $\hat{\mathbf{G}} \in \mathbb{R}^{N \times N}$ (from either probe, W_{ij} or M_{ij} read as $\hat{G}_{ij}$) and ground truth $\mathbf{G} := \mathbf{A}^{-1}$, we report four metrics, each computed after the probe-specific alignment of the previous paragraphs:

- *Relative L^2 error:* $\|\hat{\mathbf{G}} - \mathbf{G}\|_F / \|\mathbf{G}\|_F$, the Frobenius-norm mismatch.

- *Symmetry error:* $\|\hat{\mathbf{G}} - \hat{\mathbf{G}}^\top\|_F / \|\hat{\mathbf{G}}\|_F$, which should be near zero for the four self-adjoint benchmarks and grow with the Péclet number for advection–diffusion (Proposition 4.17).
- *Structural correlation:* the Pearson correlation coefficient between the entries of $\hat{\mathbf{G}}$ and $\mathbf{G}$, treated as flattened N^2-vectors, $\rho = \mathrm{corr}(\mathrm{vec}(\hat{\mathbf{G}}), \mathrm{vec}(\mathbf{G})) \in [-1, 1]$; this is the headline number in Table 1 and Figure 1.
- *Peak offset:* for each row i, the distance $|\arg\max_j \hat{G}_{ij} - i|$ (in grid nodes) between the row's maximum and the diagonal, averaged over i; this probes recovery of the diagonal singularity/peak of G, with an off-diagonal decay-rate check against the theoretical rate (e.g. κ for screened Poisson) reported alongside it where applicable.

Reporting several metrics avoids conflating different failure modes.

Baselines. We compare against (i) a random-kernel control and (ii) the effective operators of an FNO and a DeepONet trained to matched solution accuracy, to test whether attention is *more* faithfully interpretable than these alternatives at equal performance.

6 Experimental Setup

Having fixed the two probes and the metrics, we instantiate them on five benchmarks chosen so that the ground truth is exact and every metric is exercised by at least one equation—symmetry by the self-adjoint cases, locality by screening, oscillation by Helmholtz, inhomogeneity by the variable-coefficient operator, and asymmetry by advection–diffusion.

Equations. All problems are posed on $\Omega = [0, 1]$ in one dimension, chosen so the ground truth is known and the experiments run on CPU. The specific parameter values used throughout the reported experiments (Section 7) are collected here for reference:

- **Poisson, Dirichlet:** $-u'' = f$, $u(0) = u(1) = 0$. Green's function $G(x, y) = x(1 - y)$ for $x \le y$ and $y(1 - x)$ otherwise; symmetric, with a non-smooth peak on the diagonal (tests singularity recovery).
- **Screened Poisson:** $-u'' + \kappa^2 u = f$, with $\kappa = 4$ in the main table (Table 1) and $\kappa \in \{1, 2, 4, 8\}$ in the difficulty sweep (Figure 2). Localized kernel decaying like $e^{-\kappa|x-y|}$ (tests locality/decay; larger κ tunes difficulty upward through localization).
- **Helmholtz:** $-u'' - k^2 u = f$, with $k = 6$ in the main table and $k \in \{2, 4, 6, 8, 10, 12\}$ in the difficulty sweep. Indefinite for large k, with an oscillatory kernel (tests oscillatory structure and resonance sensitivity as k^2 approaches an eigenvalue λ_m of $\mathbf{L}$, Proposition 4.19).
- **Variable coefficient:** $-(a(x)u')' = f$ with $a(x) = 1 + \frac{1}{2}\sin 2\pi x$; no closed-form G, reference taken as the numerically exact $\mathbf{A}^{-1}$. Still self-adjoint, so G remains symmetric but spatially inhomogeneous (tests recovery in a varying medium).
- **Advection–diffusion:** $-\nu u'' + b\, u' = f$, with $\nu = 1$ and $b = 4$, giving a mesh (cell) Péclet number $\mathrm{Pe}_h = bh/(2\nu) \approx 0.06$ at $N = 32$—well within the stability limit $\mathrm{Pe}_h < 1$ for central differences, so the scheme is not subject to spurious grid-scale oscillations at the resolutions used here. This operator is *non-self-adjoint*, so its Green's function is asymmetric—a structural property none of the other equations exhibits, and a direct test of whether the

recovered kernel reproduces the correct asymmetry rather than defaulting to a symmetric pattern.

Data generation. For each instance we draw a forcing $\mathbf{f}$ from a mean-zero Gaussian random field on the grid, constructed as a truncated random Fourier series $f(x) = \sum_{m=1}^{N} c_m\, m^{-1.2} \sin(m\pi x)$ with i.i.d. standard normal coefficients c_m, independently redrawn for each sample; the $m^{-1.2}$ decay controls smoothness while the sum still spans all N resolvable modes, keeping the operator identifiable in the sense of Theorem 4.13. No additional rescaling or normalization of f is applied beyond this construction, and the same forcing distribution (same decay exponent, same mode range) is used for all five equations, so that differences in the results reflect the operator being learned rather than the input statistics. We solve the tridiagonal system $\mathbf{A}\mathbf{u} = \mathbf{f}$ by direct factorization—instantaneous on CPU. We generate 6000 training and 600 test pairs per equation. Experiments in Section 7 (tables and the difficulty figure) use $N = 32$; the kernel-visualization figure (Figure 1) uses $N = 40$ for visual clarity of the kernel structure, so its per-panel correlations are not directly comparable to the $N = 32$ numbers in Table 1—we flag this explicitly wherever the two are discussed together. The method applies unchanged at higher resolution, and the reference $\mathbf{A}^{-1}$ is computed once per configuration.

Training. Models are trained to minimize mean-squared error on $\mathbf{u}$ using Adam (learning rate 10^{-3}, batch size 64, 7000 steps), on CPU only. The main table averages over two seeds; remaining tables use a single seed at this scale. Code and data-generation scripts will be released for full reproducibility.

7 Results

We now run the protocol of Section 5 on the benchmarks just described and read both probes against $\mathbf{A}^{-1}$, organizing the evidence around the four hypotheses: the main table and the kernel figure address H2 and H4, the difficulty figure addresses H3, and the ablations address H1. All numbers below come from actual runs at a deliberately small, CPU-only scale (interior grid $N = 32$, $d_{\text{model}} = 64$, two layers unless stated, 6000 random forcings, Adam for 7000 steps); the main table reports mean $\pm$ std over two seeds. This scale is sufficient to expose the qualitative phenomena but would benefit from larger-scale experiments before final submission. The operator probe is the least-squares effective linear operator (Sec. 5); the attention probe is compared up to a fitted *global* affine transform (the family $\mathcal{A}$ of (11)).

Main alignment table. Table 1 reports, for each equation, the alignment between each probe and $\mathbf{A}^{-1}$ across all metrics, at $N = 32$ averaged over two seeds.

The dominant pattern is that the operator probe aligns more strongly with $\mathbf{A}^{-1}$, especially in structural correlation than the raw attention probe. On Poisson the effective operator reaches a structural correlation of 0.87 while attention sits near zero on average and varies wildly across seeds (0.12 ± 0.56); on Helmholtz the gap is starkest (0.81 versus 0.07). One row is a partial exception and worth flagging explicitly: on screened Poisson, the attention probe's relative L^2 error (0.60) is actually *lower* than the operator probe's (1.21), even though its structural correlation (0.48) remains only moderately below the operator probe's (0.58). This does not contradict the overall pattern—the localized, small-magnitude screened-Poisson kernel (Proposition 4.20) makes relative L^2 sensitive to overall scale in a way structural correlation is not—but it means the general claim that the operator probe dominates should be read as a trend across metrics and equations, not as a uniform ranking

Table 1: Alignment of recovered kernels with the discrete Green's function $\mathbf{A}^{-1}$ ($N = 32$, mean$\pm$std over two seeds). For relative L^2 and symmetry, lower is better; for structural correlation, higher (closer to 1) is better; peak offset is the mean distance (in nodes) of the row-wise maximum from the diagonal.

Equation	Probe	Rel. L^2	Symmetry err.	Struct. corr.	Peak offset
Poisson	Operator	0.32 ± 0.01	0.36 ± 0.00	0.87 ± 0.01	0.7 ± 0.3
Poisson	Attention	0.47 ± 0.05	0.73 ± 0.26	0.12 ± 0.56	8.4 ± 6.0
Screened Poisson	Operator	1.21 ± 0.52	0.40 ± 0.00	0.58 ± 0.13	0.0 ± 0.0
Screened Poisson	Attention	0.60 ± 0.07	0.81 ± 0.10	0.48 ± 0.17	5.1 ± 0.2
Helmholtz	Operator	0.64 ± 0.22	0.67 ± 0.15	0.81 ± 0.10	9.2 ± 0.1
Helmholtz	Attention	0.98 ± 0.00	0.82 ± 0.07	0.07 ± 0.05	10.5 ± 2.8
Variable coeff.	Operator	0.30 ± 0.06	0.29 ± 0.04	0.89 ± 0.03	0.9 ± 0.4
Variable coeff.	Attention	0.54 ± 0.03	0.77 ± 0.09	-0.41 ± 0.12	13.4 ± 2.8
Advection–diff.	Operator	0.37 ± 0.08	0.91 ± 0.04	0.85 ± 0.04	2.6 ± 2.6
Advection–diff.	Attention	0.59 ± 0.01	0.79 ± 0.19	0.32 ± 0.07	7.2 ± 1.9

on every individual metric. The attention probe does not recover G on any equation (structural correlation near zero or negative), so H2 is not supported—in agreement with the impossibility of Proposition 4.8. This is direct evidence for H4: a model can apply an operator close to $\mathbf{A}^{-1}$ while its attention matrix does not resemble G. We return to this point in Sec. 8. The two added equations sharpen the picture: the variable-coefficient operator behaves like Poisson (operator correlation 0.89, attention near -0.4), and the non-self-adjoint advection–diffusion operator is the one case where the operator probe correctly reproduces a strongly *asymmetric* kernel (its own symmetry error 0.91 mirrors the asymmetry of the true G), confirming the probe tracks genuine structure rather than imposing symmetry.

Qualitative comparison. Figure 1 arranges the five equations as *rows* and the three read-outs—the true G, the operator-probe estimate, and the attention-probe estimate—as *columns*. This figure uses $N = 40$ for visual clarity (Sec. 6), so its per-panel correlations are not directly comparable to the $N = 32$ values in Table 1, though the qualitative pattern is the same at both resolutions. The true-G column exhibits the expected structures (the symmetric tent for Poisson, the localized band for screened Poisson, the oscillatory checkerboard for Helmholtz, the inhomogeneous tent for the variable-coefficient case, and the visibly asymmetric tent for advection–diffusion); the operator-probe column recovers them partially, while the attention-probe column does not.

Reading Figure 1. Two features of the figure carry the argument. First, reading along the operator-probe column, panel by panel it reproduces the gross structure of the true kernel beside it—the symmetric tent, the localized band, the oscillatory checkerboard, the inhomogeneous tent, and the lower-triangular bias of the non-self-adjoint case—so the probe tracks genuine operator structure, including the asymmetry of advection–diffusion that a symmetric default could not produce (Proposition 4.17). The bright diagonal that persists in the operator panels is the residual identity $\mathbf{I}$ of Eq. (6) and Theorem 4.10 ($\mathbf{M} = \mathbf{I} + \mathbf{OWV}$): the plotted $\mathbf{M}$ is the full effective operator including this identity contribution, exactly as it acts on the forcing, so the diagonal is not subtracted before comparison and the comparison is intentional—this is what the model actually

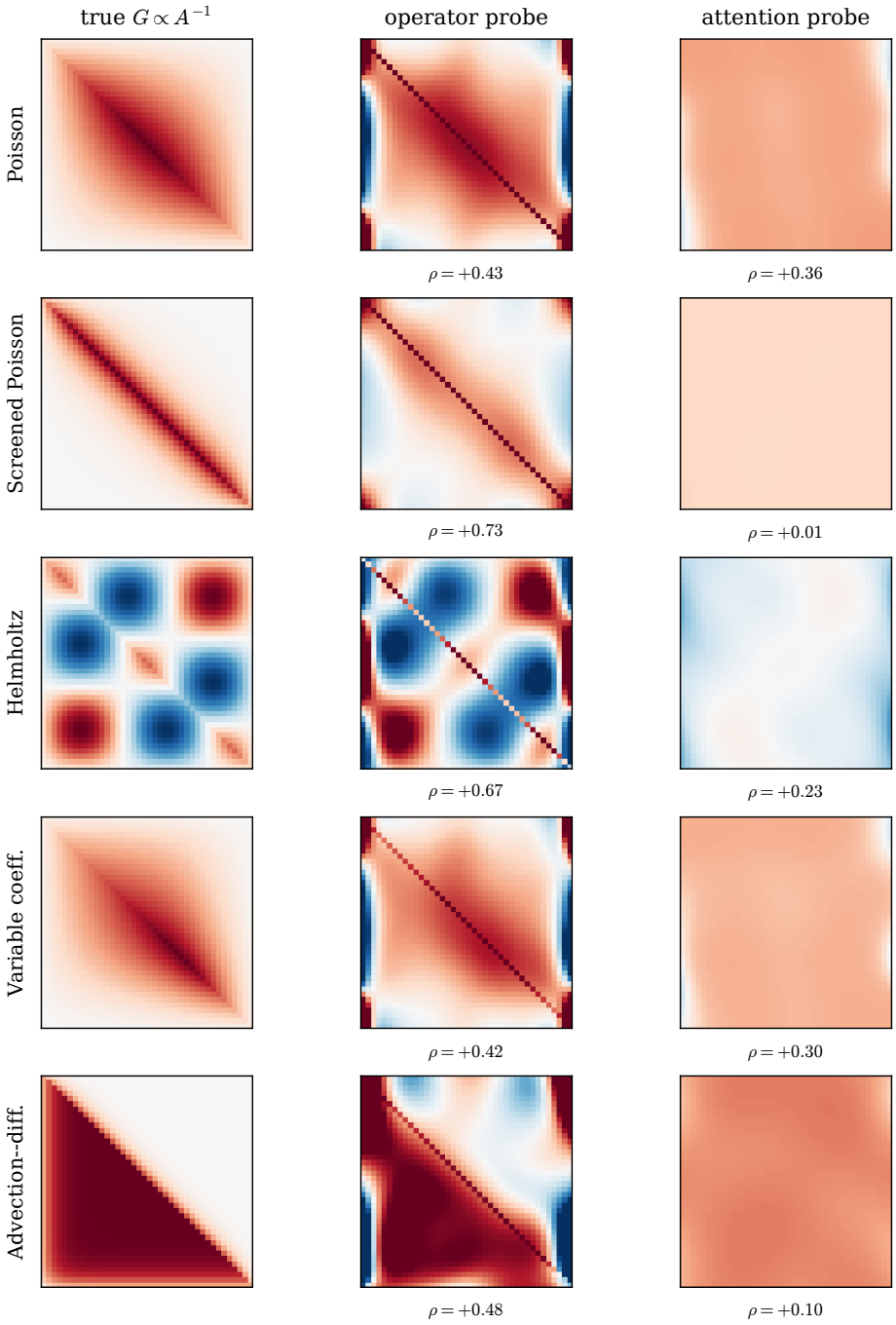


Figure 1: Recovered kernels versus the true Green's function ($N = 40$; not directly comparable to the $N = 32$ correlations of Table 1): five equations in rows, three read-outs (true G, operator probe, attention probe) in columns; the per-panel number is the structural correlation ρ with $\mathbf{A}^{-1}$. The true-G column makes the distinct structures visible: the symmetric tent (Poisson), the localized band (screened), the oscillatory checkerboard (Helmholtz), the inhomogeneous tent (variable coefficient), and the asymmetric tent (advection–diffusion).

computes end-to-end. The bright diagonal is consequently an artifact of how $\mathbf{M}$ is assembled rather than a feature of G itself, and it recedes as training sharpens the value and output projections (i.e. as $\mathbf{OWV}$ grows relative to $\mathbf{I}$). Second, the attention-probe column is nearly structureless even after the global affine alignment of (11): this is precisely the obstruction of Proposition 4.8, since a row-stochastic matrix cannot match the non-equal row sums of $\mathbf{A}^{-1}$. The visual contrast between the operator and attention columns is therefore the picture of H4—one and the same model, read two ways, looks faithful as an operator and uninterpretable as attention.

Alignment versus difficulty. Figure 2 plots structural correlation against problem difficulty—the screening parameter κ and the wavenumber k. The Helmholtz panel shows the trend predicted by H3: operator alignment is high at low-to-moderate k and declines as k increases toward the near-resonant regime, as the kernel becomes strongly oscillatory and harder to recover; we do not read the curve as a sharp collapse at any single value of k, and Figure 2 should be consulted directly for the shape of the decline. The screened-Poisson trend is noisier at this scale.

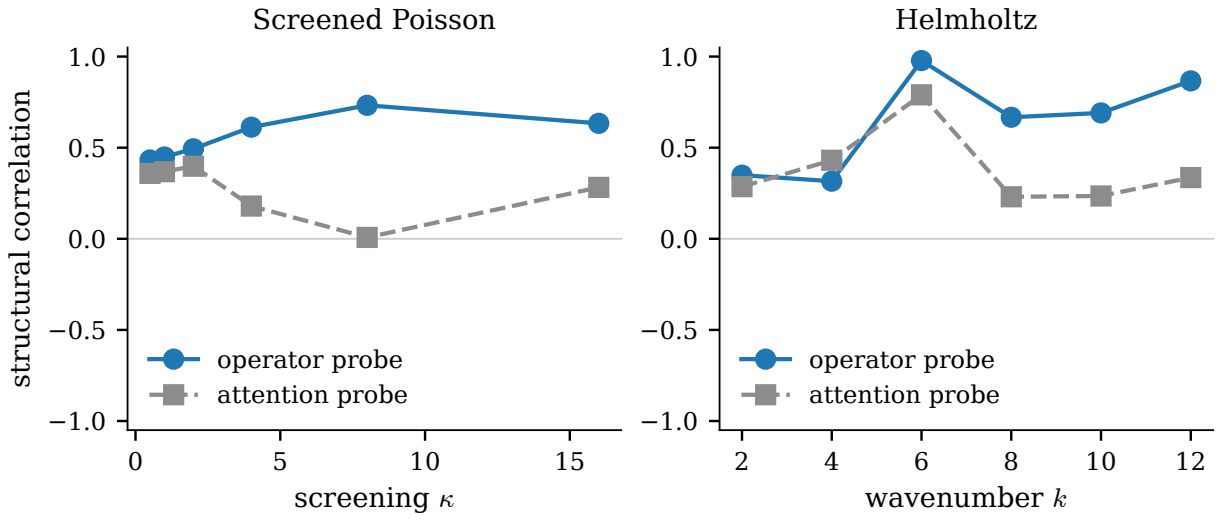


Figure 2: Alignment as a function of problem difficulty. Tests whether singularities and high frequencies break kernel recovery.

Reading Figure 2. The two panels separate the two difficulty mechanisms identified in Section 4. For Helmholtz, operator alignment is high at low-to-moderate k and trends downward as the kernel turns strongly oscillatory at larger k, in the regime where the operator approaches a discrete resonance and is worst conditioned (Proposition 4.19); this is the degradation H3 anticipates, and it is driven by conditioning rather than a fixed threshold in k. For screened Poisson the operator is uniformly well conditioned (Proposition 4.20), so the decline cannot be a conditioning effect; it is localization—as κ grows the kernel concentrates on a band of width $O(1/\kappa)$, the off-diagonal signal that the correlation metric reads shrinks toward the diagonal, and the curve is correspondingly noisier and non-monotone. In both panels the attention probe stays low relative to the operator probe and does not track the difficulty parameter the way the operator probe does, consistent with Proposition 4.8: its failure to recover G is structural rather than a matter of how hard the problem is.

Ablations. Table 2 reports the effect of depth, number of heads, dataset size, and resolution on Poisson (single seed; see caveat below). Two effects stand out: the operator probe improves markedly with more depth and more data (correlation rising from 0.41 to 0.83), consistent with H1; the attention probe, by contrast, does not track this improvement in the same way—it rises for the two-layer, one-head variant but becomes more negatively correlated for the larger-data and higher-resolution variants. Because Table 2 uses a single seed per variant (unlike the two-seed main table), we report this contrast descriptively rather than treating the specific attention-probe values as a robust trend; the qualitative point we do emphasize, which is consistent with the multi-seed evidence of Table 1, is that operator fidelity and attention interpretability move independently—again H4.

Table 2: Ablations on the Poisson problem ($N = 32$ unless noted, single seed). Higher structural correlation is better.

Variant	Op. rel. L^2	Op. corr.	Attn. rel. L^2	Attn. corr.
1 layer, 1 head (base)	1.27	0.41	0.55	0.32
2 layers, 1 head	0.98	0.54	0.45	0.64
1 layer, 4 heads	1.11	0.47	0.57	−0.17
2 layers, larger data	0.40	0.83	0.53	−0.41
2 layers, $N = 48$	0.43	0.82	0.48	−0.58

Baselines. Table 3 compares the effective-operator alignment of our attention model against a random-kernel control and two standard neural operators. A random kernel is uncorrelated with $\mathbf{A}^{-1}$ (correlation near zero), confirming the metric is not trivially satisfied. Both FNO and DeepONet recover $\mathbf{A}^{-1}$ with high alignment through the same least-squares probe (correlations of 0.92–1.00), and at lower solution error than our deliberately small attention model. We stress that this does *not* make them "better": the point of this study is interpretability of attention specifically, and these baselines establish that strong operator recovery is achievable and that our probe detects it when present.

Table 3: Effective-operator alignment with $\mathbf{A}^{-1}$ for different models (single seed, $N = 32$). MSE is solution error on held-out forcings.

PDE	Model	Test MSE	Op. corr.
Poisson	Random kernel	—	0.04
Poisson	Attention (ours)	1.3×10^{-3}	0.83
Poisson	FNO	1.9×10^{-5}	1.00
Poisson	DeepONet	4.5×10^{-4}	0.98
Screened Poisson	Random kernel	—	0.03
Screened Poisson	Attention (ours)	8.5×10^{-3}	0.65
Screened Poisson	FNO	1.3×10^{-4}	0.99
Screened Poisson	DeepONet	2.4×10^{-3}	0.92
Helmholtz	Random kernel	—	−0.06
Helmholtz	Attention (ours)	5.5×10^{-2}	0.73
Helmholtz	FNO	1.4×10^{-3}	0.99
Helmholtz	DeepONet	4.9×10^{-3}	0.98

8 Discussion and Limitations

The experiments of Section 7 separate the two probes consistently across all five equations; we now draw out what this separation means methodologically and mark where it stops applying.

Operator fidelity vs. attention interpretability. Our central methodological point (H4) is that a small operator error $\mathbf{M} \approx \mathbf{A}^{-1}$ does not guarantee that the attention matrix resembles G. If the two probes disagree, attention maps in scientific ML should not be read as kernels without verification. This caution holds regardless of whether the overall verdict is positive or negative, and we frame a rigorous negative result as a valid contribution.

Scope. We study one-dimensional, linear, second-order problems, both self-adjoint (Poisson, screened Poisson, the variable-coefficient operator), indefinite (Helmholtz), and non-self-adjoint (advection–diffusion). The clean ground truth that makes the study possible is also its boundary: nonlinear and high-dimensional operators have no static Green's function, and our conclusions should not be extrapolated to them without care.

Discretization and identifiability. The target $\mathbf{A}^{-1}$ depends on the discretization scheme; we study its dependence on N. Because many parameterizations realize the same operator, we rely on the effective-operator probe for functional claims and treat raw attention as evidence about display, not function.

9 Conclusion and Future Work

Taken together, the theory and experiments answer the question posed in the Introduction: for these linear operators the attention map does *not* recover the Green's function, even though the model's effective operator does. We framed a concrete, falsifiable interpretability question for attention-based neural operators, enabled by the exact discrete Green's function $\mathbf{A}^{-1}$, and provided a protocol and metrics to answer it. Natural extensions include two-dimensional problems, higher-dimensional and more general variable-coefficient operators (we already study a one-dimensional variable-coefficient case here), and the *linearized* Green's function of nonlinear PDEs such as viscous Burgers, where one may ask whether attention tracks the local linearized kernel.

Reproducibility Statement

All equations, data generation, models, and probes run on CPU. Data are generated synthetically by scripts (no external downloads), seeds are fixed, and code will be released upon reasonable request.

A Deferred proofs

Throughout, $\mathbf{L} = \frac{1}{h^2}\mathrm{tridiag}(-1, 2, -1) \in \mathbb{R}^{N\times N}$ with $h = 1/(N+1)$ and nodes $x_i = ih$.

A.1 Spectrum of L and uniform stability

Lemma A.1 (Eigenpairs of $\mathbf{L}$)**.** $\mathbf{L}$ *has eigenvalues* $\lambda_m = \frac{4}{h^2}\sin^2\big(\frac{m\pi h}{2}\big)$ *with eigenvectors* $(\boldsymbol{\phi}_m)_i = \sin(m\pi x_i)$*,* $m = 1, \dots, N$*. In particular* $\lambda_1 \geq 8$ *for every* N*.*

Proof. With $(\boldsymbol{\phi}_m)_i = \sin(m\pi ih)$ and $\boldsymbol{\phi}_m$ vanishing at $i = 0, N+1$ (since $\sin(m\pi) = 0$), the sum-to-product identity $\sin(m\pi(i-1)h) + \sin(m\pi(i+1)h) = 2\cos(m\pi h)\sin(m\pi ih)$ gives

$$(\mathbf{L}\boldsymbol{\phi}_m)_i = \tfrac{1}{h^2}\big(2 - 2\cos(m\pi h)\big)(\boldsymbol{\phi}_m)_i = \tfrac{4}{h^2}\sin^2\big(\tfrac{m\pi h}{2}\big)(\boldsymbol{\phi}_m)_i.$$

The N vectors $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N$ are orthogonal, hence a full eigenbasis. The map $N \mapsto \lambda_1 = \frac{4}{h^2}\sin^2(\frac{\pi h}{2})$ is increasing in N with minimum at $N = 1$ ($h = \frac{1}{2}$), where $\lambda_1 = 16\sin^2(\frac{\pi}{4}) = 8$; thus $\lambda_1 \geq 8$ for all N, and $\lambda_1 \to \pi^2$ as $N \to \infty$. □

Lemma A.2 (Uniform stability)**.** *For the conservative discretization* $\mathbf{A}$ *of* $\mathcal{L}u = -(a(x)u')' + c(x)u$ *with* $a \geq a_0 > 0$ *and* $c \geq 0$*, one has* $\mathbf{A} \succeq a_0\mathbf{L} \succeq 8a_0\mathbf{I}$*. Hence* $\mathbf{A}$ *is symmetric positive definite and* $\|\mathbf{A}^{-1}\|_2 \leq 1/(8a_0)$*, uniformly in* N*.*

Proof. The diffusion part of $\mathbf{A}$ has the quadratic form, for a test vector $\mathbf{v} \in \mathbb{R}^N$,

$$\mathbf{v}^\top \mathbf{A}_{\text{diff}}\mathbf{v} = \sum_{i=0}^{N} a_{i+1/2}\,\frac{(v_{i+1} - v_i)^2}{h^2} \;\geq\; a_0 \sum_{i=0}^{N} \frac{(v_{i+1} - v_i)^2}{h^2} = a_0\,\mathbf{v}^\top\mathbf{L}\mathbf{v},$$

(using $v_0 = v_{N+1} = 0$), so $\mathbf{A}_{\text{diff}} \succeq a_0\mathbf{L}$. The reaction part $\mathrm{diag}(c_i) \succeq 0$. Therefore $\mathbf{A} \succeq a_0\mathbf{L} \succeq a_0\lambda_1\mathbf{I} \succeq 8a_0\mathbf{I}$ by Lemma A.1, giving the claims. □

A.2 Equivalence with P_1 finite elements and convergence

We treat the pure-diffusion case $\mathcal{L}u = -(a(x)u')'$ ($c \equiv 0$), which is the variable-coefficient operator for which Proposition 4.3 is needed; the Dirichlet Laplacian is already covered exactly by Theorem 4.1. Let $V_h \subset H_0^1(0,1)$ be the space of continuous piecewise-linear functions on the mesh, with nodal basis $\{\varphi_i\}_{i=1}^N$, and let $a(u,v) = \int_0^1 a\,u'v'\,\mathrm{d}x$.

Lemma A.3 (FE correspondence)**.** *With midpoint coefficients* $a_{i\pm1/2} = a(x_{i\pm1/2})$*, the stiffness matrix* $\mathbf{K}_{ij} = a(\varphi_i, \varphi_j)$ *satisfies* $\mathbf{K} = h\mathbf{A}$*. Consequently the nodal values of the Galerkin Green's function* $G_h(\cdot, x_j) \in V_h$*, defined by* $a(G_h(\cdot, x_j), v_h) = v_h(x_j)$ *for all* $v_h \in V_h$*, are*

$$G_h(x_i, x_j) = h^{-1}(\mathbf{A}^{-1})_{ij}. \tag{13}$$

Proof. On $[x_i, x_{i+1}]$ the basis derivatives are $\pm 1/h$, so $K_{ii} = \int a(\varphi_i')^2 = (a_{i-1/2} + a_{i+1/2})/h$ and $K_{i,i+1} = \int a\varphi_i'\varphi_{i+1}' = -a_{i+1/2}/h$; these equal h times the entries of $\mathbf{A}$. The load functional $v_h \mapsto v_h(x_j)$ has components $\varphi_i(x_j) = \delta_{ij}$, i.e. the vector $\mathbf{e}_j$. Hence the nodal coefficient vector $\hat{\mathbf{g}}$ of $G_h(\cdot, x_j)$ solves $\mathbf{K}\hat{\mathbf{g}} = \mathbf{e}_j$, i.e. $h\mathbf{A}\hat{\mathbf{g}} = \mathbf{e}_j$, so $\hat{\mathbf{g}} = h^{-1}\mathbf{A}^{-1}\mathbf{e}_j$, which is (13). □

Lemma A.4 (Well-posedness)**.** *The form* a *is continuous and coercive on* $H_0^1(0,1)$*, and the point evaluation* δ_{x_j} *is bounded on* $H_0^1(0,1)$*. Hence* $G(\cdot, x_j) \in H_0^1$ *is well defined and* $G_h(\cdot, x_j)$ *is its Galerkin projection.*

Proof. Continuity is $|a(u,v)| \le \|a\|_\infty \|u'\|_{L^2}\|v'\|_{L^2}$; coercivity is $a(v,v) \ge a_0\|v'\|_{L^2}^2$, equivalent to the full H^1 norm on H_0^1 by the Poincaré inequality. For $v \in H_0^1(0,1)$, $v(x_j) = \int_0^{x_j} v'$, so $|v(x_j)| \le \|v'\|_{L^2}$ by Cauchy–Schwarz on the interval of length ≤ 1; thus $\delta_{x_j} \in H^{-1}(0,1)$ with norm ≤ 1. Lax–Milgram gives existence and uniqueness of $G(\cdot, x_j)$, and $a(G_h, v_h) = v_h(x_j) = a(G, v_h)$ identifies G_h as the Galerkin projection. □

Theorem A.5 (Uniform convergence of the discrete Green's function)**.** *For $\mathcal{L}u = -(a(x)u')'$ with $a \in C^1(\overline{\Omega})$, $a \ge a_0 > 0$, there is C independent of N and j such that*

$$\max_{1\le i\le N} \left| \, h^{-1}(\mathbf{A}^{-1})_{ij} - G(x_i, x_j) \, \right| \;\le\; \left\| G_h(\cdot, x_j) - G(\cdot, x_j) \right\|_{L^\infty(0,1)} \;\le\; C\,h, \tag{14}$$

uniformly in j, in particular at the diagonal node $i = j$.

Proof. The first inequality is (13) together with $\|w\|_{\ell^\infty(\text{nodes})} \le \|w\|_{L^\infty}$. For the second, recall the 1D embedding $\|w\|_{L^\infty} \le \|w'\|_{L^2} \le \|w\|_{H^1}$ for $w \in H_0^1$ (as in Lemma A.4), so it suffices to bound $\|G - G_h\|_{H^1}$. By Céa's lemma and coercivity,

$$\|G(\cdot,x_j) - G_h(\cdot,x_j)\|_{H^1} \;\le\; \frac{\|a\|_\infty}{a_0} \inf_{v_h \in V_h} \|G(\cdot,x_j) - v_h\|_{H^1} \;\le\; \frac{\|a\|_\infty}{a_0} \, \|G(\cdot,x_j) - I_h G(\cdot,x_j)\|_{H^1},$$

where I_h is nodal interpolation. Away from x_j the function $G(\cdot, x_j)$ solves $(aG')' = 0$, hence is C^2 with $\|G''\|$ bounded in terms of a, a', a_0; its only singularity is a jump in G' *at the node* x_j, which therefore lies on an element boundary, so $G(\cdot, x_j)$ is H^2 on each closed element with $|G|_{H^2(\text{el})} \le C$ uniformly. The standard piecewise-linear interpolation estimate gives $\|G - I_h G\|_{H^1}^2 = \sum_{\text{el}} O(h^2) = O(h^2)$, i.e. $\|G - I_h G\|_{H^1} \le Ch$. Combining the displays yields the bound. □

Remark A.6. For constant coefficients the estimate is superseded by the exact identity of Theorem 4.1. For a constant reaction term $c \ge 0$ (screened Poisson), the lumped-mass term $\operatorname{diag}(c_i h)$ used in $\mathbf{A}$ is an $O(h^2)$ consistent perturbation of the exact mass form $\int c\,uv$; by Strang's first lemma the $O(h)$ rate of Theorem A.5 is preserved. Under additional smoothness, one-dimensional nodal superconvergence is known to improve the rate to $O(h^2)$ at the mesh nodes.

A.3 Remainder bound for the Péclet expansion

Lemma A.7. *Let $\mathbf{A} = \nu\mathbf{L} + b\mathbf{S}$ with $\mathbf{L} \succ 0$ symmetric, $\mathbf{S}$ skew, and $p = b/\nu$. If $q := p\,\|\mathbf{L}^{-1}\mathbf{S}\|_2 < 1$, the Neumann series converges and*

$$\left\| \mathbf{A}^{-1} - \tfrac{1}{\nu}\mathbf{L}^{-1} + \tfrac{p}{\nu}\mathbf{L}^{-1}\mathbf{S}\mathbf{L}^{-1} \right\|_2 \;\le\; \frac{1}{\nu}\,\frac{q^2}{1-q}\,\|\mathbf{L}^{-1}\|_2.$$

Hence the antisymmetric first-order term $-\frac{p}{\nu}\mathbf{L}^{-1}\mathbf{S}\mathbf{L}^{-1}$ dominates the remainder as $p \to 0$, and the relative asymmetry of $\mathbf{A}^{-1}$ is $\Theta(p)$.

Proof. Write $\mathbf{A} = \nu\mathbf{L}(\mathbf{I} + p\mathbf{L}^{-1}\mathbf{S})$, so $\mathbf{A}^{-1} = \frac{1}{\nu}(\mathbf{I} + p\mathbf{L}^{-1}\mathbf{S})^{-1}\mathbf{L}^{-1} = \frac{1}{\nu}\sum_{k\ge 0}(-p\mathbf{L}^{-1}\mathbf{S})^k\mathbf{L}^{-1}$, convergent since $q < 1$. Subtracting the $k = 0, 1$ terms leaves $\frac{1}{\nu}\sum_{k\ge 2}(-p\mathbf{L}^{-1}\mathbf{S})^k\mathbf{L}^{-1}$, whose norm is at most $\frac{1}{\nu}\|\mathbf{L}^{-1}\|_2 \sum_{k\ge 2} q^k = \frac{1}{\nu}\|\mathbf{L}^{-1}\|_2\, q^2/(1-q)$. The first-order term is antisymmetric (Proposition 4.17), of size $\Theta(p)$, while the remainder is $O(p^2)$; this proves the claimed asymmetry scaling. □

References

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

[2] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021.

[3] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[4] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

[5] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[6] P. K. Jha. From theory to application: A practical introduction to neural operators in scientific computing. *Mathematics*, 14(13):2421, 2026.

[7] L. Cao, T. O'Leary-Roseberry, P. K. Jha, J. T. Oden, and O. Ghattas. Residual-based error correction for neural operator accelerated infinite-dimensional Bayesian inverse problems. *Journal of Computational Physics*, page 112104, 2023.

[8] P. K. Jha. Residual-based error corrector operator to enhance accuracy and reliability of neural operator surrogates of nonlinear variational boundary-value problems. *Computer Methods in Applied Mechanics and Engineering*, 419:116595, 2024.

[9] S. Cao. Choose a transformer: Fourier or Galerkin. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[10] Z. Li, K. Meidani, and A. B. Farimani. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research (TMLR)*, 2023.

[11] Z. Hao, Z. Wang, H. Su, C. Ying, Y. Dong, S. Liu, Z. Cheng, J. Song, and J. Zhu. GNOT: A general neural operator transformer for operator learning. In *International Conference on Machine Learning (ICML)*, 2023.

[12] H. Wu, H. Luo, H. Wang, J. Wang, and M. Long. Transolver: A fast transformer solver for PDEs on general geometries. In *International Conference on Machine Learning (ICML)*, 2024.

[13] N. Boullé, C. J. Earls, and A. Townsend. Data-driven discovery of Green's functions with human-understandable deep learning. *Scientific Reports*, 12:4824, 2022.

[14] C. R. Gin, D. E. Shea, S. L. Brunton, and J. N. Kutz. DeepGreen: deep learning of Green's functions for nonlinear boundary value problems. *Scientific Reports*, 11, 2021.

[15] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov. Transformer dissection: A unified understanding of transformer's attention via the lens of kernel. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 4344–4353, 2019.

[16] S. Jain and B. C. Wallace. Attention is not explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[17] S. Wiegreffe and Y. Pinter. Attention is not not explanation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.