10.48047/jocaaa.2023.31.04.66

From PageRank to Transformers: Tracing the Evolution of Attention and Ranking Mechanisms in Modern AI

Ravi Teja Gurram
Independent Researcher, Orlando, Florida, United States
ORCID: 0009-0008-4406-8126
Email: gurramraviteja1@gmail.com

Abstract

Two of the most consequential ideas in modern computing—PageRank, which ordered the early web, and the attention mechanism, which powers today's large language models—are usually told as separate stories. This article argues they are chapters of one story about computing importance over a set of related elements. PageRank assigns each web page a score equal to the probability that an infinite random walk, with occasional teleportation, lands on it; mathematically this score is the stationary distribution of a Markov chain, equivalently the dominant eigenvector of a stochastic matrix, found by power iteration. Attention assigns each element of a sequence a weight reflecting its relevance to a query, computed as a softmax over compatibility scores and used to form a weighted average of value vectors. We trace the conceptual line connecting these mechanisms—through graph attention networks, which apply attention over a node's neighbors, to approaches such as APPNP that derive neural message passing directly from personalized PageRank—and we make the shared structure precise: both are linear operators that normalize a set of nonnegative affinities into a weighting and aggregate accordingly. We support the synthesis with a small, fully reproducible study. It verifies that PageRank computed by power iteration and by eigen-decomposition agree to within 10^{-13}, that convergence slows predictably as the damping factor rises, and that a softmax attention distribution and a PageRank transition row are the same kind of object—normalized weights over a set, differing only in whether those weights are fixed by graph structure or learned from data. We close with a synthesis and a research agenda on unifying ranking and attention.

Keywords: PageRank; attention; Transformers; Markov chains; eigenvectors; graph neural networks; personalized PageRank; ranking

1. Introduction

In 1998, two graduate students proposed that the importance of a web page could be measured by the behavior of an idealized random surfer: a page is important if a surfer following links at random spends a large fraction of time there. This idea, PageRank [1], became the ranking backbone of a search engine that reshaped the internet. Nearly two decades later, a different idea reshaped artificial intelligence: that a model could decide, for each element of an input, how much to attend to every other element, and that this attention was, in the authors' words, all you need [2]. These two ideas are usually presented as belonging to different worlds—information retrieval and deep learning. This article argues they are deeply related expressions of a single computational pattern.

The pattern is this: given a set of elements with pairwise affinities, compute for each element a normalized, nonnegative weighting over the others, and use that weighting to produce an importance score or an aggregated representation. PageRank computes a global importance vector by iterating a normalized link

10.48047/jocaaa.2023.31.04.66

matrix to its stationary distribution. Attention computes, for each query, a local weighting over keys via a softmax of compatibility scores, and aggregates the corresponding values. Both are, at heart, the repeated or one-shot application of a row-stochastic operator to propagate importance through a set of relationships.

This relationship is not merely an analogy. It has been made formal and exploited in practice. Graph attention networks [3] apply an attention mechanism over a node's graph neighbors, directly fusing the ranking-over-a-graph idea with the learned-weighting idea. The APPNP model [4] goes further, deriving its propagation rule from personalized PageRank and showing that a neural network's predictions can be diffused across a graph exactly as importance diffuses in a random walk with restart. Reading PageRank and attention together, with these bridges in view, clarifies both and suggests where they might converge further.

1.1 Contributions

1. A unifying account of PageRank and attention as instances of one pattern: normalized weighting and aggregation over a set of related elements.
2. A clear exposition of the mathematical bridges—graph attention networks and personalized-PageRank-based propagation—that formally connect ranking and attention.
3. A reproducible study verifying the equivalence of PageRank's two computations, the effect of damping on convergence, and the shared structure of attention weights and transition rows.
4. A synthesis and research agenda on unified ranking-attention mechanisms.

1.2 Organization

Section 2 states the problem and gap. Section 3 gives objectives and questions. Section 4 reviews PageRank, attention, and their bridges. Section 5 formalizes the shared mathematics. Section 6 describes the empirical methodology; Sections 7 and 8 report and discuss results. Sections 9 through 12 cover implications, limitations, future scope, and conclusions.

2. Problem Statement and Research Gap

2.1 Problem Statement

Students and practitioners typically learn PageRank in a course on information retrieval or linear algebra and attention in a course on deep learning, with no bridge between them. This separation obscures transferable intuition: that the stability and convergence theory of Markov chains illuminates the behavior of iterative attention-like propagation, and that the learned, query-dependent weighting of attention generalizes the fixed, structure-dependent weighting of PageRank. A unified treatment would let insights flow both ways.

2.2 Research Gap

While individual bridges exist in the research literature—graph attention [3] and personalized-PageRank propagation [4] are well known to specialists—there is little accessible synthesis that presents PageRank and attention as one lineage and demonstrates the shared structure concretely. The gap is expository and integrative: a treatment that states the common pattern, traces the historical and mathematical links, and verifies the connection with reproducible computation. This article addresses that gap.

10.48047/jocaaa.2023.31.04.66

Table 1. *The lineage of mechanisms reviewed, framed as variations on weighted aggregation over a set.*

Mechanism (year)	Elements	Weights come from	Aggregation
PageRank (1998) [1]	Web pages	Link structure (fixed)	Stationary distribution
Attention (2017) [2]	Sequence tokens	Learned query·key (dynamic)	Weighted sum of values
GAT (2018) [3]	Graph nodes	Learned over neighbors	Weighted sum of neighbors
APPNP (2019) [4]	Graph nodes	Personalized PageRank	PPR-diffused predictions

3. Objectives and Research Questions

3.1 Objectives

5. To present PageRank and attention under a single weighted-aggregation framework.
6. To explain the mathematical bridges that formally connect ranking and attention.
7. To verify the connections reproducibly: PageRank's two computations, damping and convergence, and the shared structure of weights.
8. To outline research directions toward unified ranking-attention mechanisms.

3.2 Research Questions

RQ1. In what precise sense are PageRank and attention the same kind of computation, and how do they differ?

RQ2. How does the damping (teleportation) parameter govern the convergence of PageRank's power iteration?

RQ3. Do the mathematical bridges (graph attention, personalized-PageRank propagation) reflect a genuine structural unity rather than a loose analogy?

4. Literature Review

4.1 PageRank: Ranking by Random Walk

PageRank [1] models a random surfer who, at each step, follows an outgoing link from the current page chosen uniformly at random. The importance of a page is the long-run fraction of time the surfer spends on it—the stationary distribution of the resulting Markov chain. Two complications must be handled. Pages with no outgoing links (dangling nodes) act as absorbing states, and disconnected regions prevent the walk from reaching all pages. Brin and Page resolved both with a damping factor: with probability α (classically 0.85) the surfer follows a link, and with probability 1−α it teleports to a page chosen uniformly at random. This teleportation makes the chain irreducible and aperiodic, guaranteeing by the Perron-Frobenius theorem a unique positive stationary distribution [5], [6]. In practice the distribution is computed by power iteration—repeatedly multiplying a probability vector by the transition matrix until it converges—because explicit eigendecomposition is infeasible at web scale, and because the convergence rate is governed by the damping factor, with the error after t steps bounded essentially by α raised to the t [6].

10.48047/jocaaa.2023.31.04.66

The deeper literature on PageRank connects it firmly to Markov chain theory: the Google matrix is row-stochastic with spectral radius one, its stationary vector is the dominant left eigenvector, and the teleportation term controls both uniqueness and mixing speed [6]. Personalized (or topic-sensitive) PageRank replaces the uniform teleport distribution with one concentrated on a set of seed pages, yielding importance relative to those seeds—a random walk with restart—which will prove central to the bridge with attention [4].

4.2 Attention: Weighting by Learned Compatibility

Attention mechanisms were introduced to let sequence models focus selectively on relevant inputs, originally augmenting recurrent encoder-decoders. The Transformer [2] dispensed with recurrence entirely, relying on self-attention. Each element is projected into a query, a key, and a value. The compatibility of a query with each key is measured by a scaled dot product; a softmax over these scores yields a distribution of attention weights; and the output is the weighted sum of the value vectors. The scaling by the inverse square root of the key dimension prevents the dot products from growing so large that the softmax saturates into near-zero gradients [2], [7]. Multiple attention heads run in parallel, letting the model attend to different relational patterns simultaneously. This architecture underlies essentially all modern large language models.

The crucial observation for this article is structural: the vector of attention weights for a given query is nonnegative and sums to one—it is a probability distribution over the elements, exactly like a row of a Markov transition matrix. The output is a convex combination of values under that distribution, exactly as a single step of a random walk pushes probability mass along a transition row.

4.3 The Bridges: Graph Attention and Personalized PageRank

The two mechanisms meet explicitly on graphs. Graph attention networks [3] compute, for each node, learned attention coefficients over its neighbors and update the node as the attention-weighted sum of neighbor features, normalized by a softmax—attention applied to graph structure rather than to a sequence. Independent analysis has examined exactly how expressive these coefficients are [8]. The most direct bridge to PageRank is APPNP [4], which observes that propagating predictions across a graph by repeated neighbor averaging is a random walk, and that using a personalized-PageRank diffusion instead—predict node labels from features, then propagate them by an adapted personalized PageRank—yields a larger, tunable neighborhood and avoids the over-smoothing of naive propagation. APPNP's propagation is literally a power iteration of a topic-sensitive PageRank operator [4]. Thus the random-walk mathematics of PageRank reappears, unchanged in form, inside a modern neural architecture.

Table 2. *Key works connecting ranking and attention.*

Work (year)	Contribution	Bridge to the thesis
Brin & Page (1998) [1]	PageRank	Importance via random walk
Vaswani et al. (2017) [2]	Transformer / self-attention	Importance via learned weights
Veličković et al. (2018) [3]	Graph attention networks	Attention over graph neighbors
Klicpera et al. (2019) [4]	APPNP	Propagation = personalized PageRank
Langville & Meyer (2004) [6]	PageRank theory	Markov / eigenvector foundations

10.48047/jocaaa.2023.31.04.66

5. The Shared Mathematics

5.1 PageRank as a Stationary Distribution

Let the web graph have n pages with row-stochastic transition matrix P, where entry P_{ij} is the probability of moving from page i to page j. The Google matrix adds teleportation with damping α and uniform teleport vector, where 1 is the all-ones vector.

$$G = \alpha P + (1-\alpha)\,\frac{1}{n}\,\mathbf{1}\mathbf{1}^{\top} \tag{1}$$

The PageRank vector π is the stationary distribution: the left eigenvector of G with eigenvalue 1, equivalently the fixed point of repeated multiplication.

$$\pi^{\top} = \pi^{\top} G, \qquad \sum_i \pi_i = 1 \tag{2}$$

It is computed by power iteration, applying G to a probability vector until convergence; by Perron-Frobenius the iteration converges to the unique π for α < 1, with error contracting at rate α.

$$\pi^{(t+1)\top} = \pi^{(t)\top} G, \qquad \| \pi^{(t)} - \pi \| \leq C\,\alpha^{t} \tag{3}$$

5.2 Attention as Normalized Weighting

Given queries Q, keys K, and values V (rows are elements), scaled dot-product attention computes a weight matrix by softmax over compatibilities and aggregates the values, where d_k is the key dimension.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{4}$$

For a single query q, the weight assigned to element j is a softmax over scores, and the output is the weighted sum of values—a convex combination, since the weights are nonnegative and sum to one.

$$\alpha_j = \frac{\exp(q\cdot k_j/\sqrt{d_k})}{\sum_l \exp(q\cdot k_l/\sqrt{d_k})}, \qquad \text{out} = \sum_j \alpha_j\, v_j \tag{5}$$

5.3 The Unifying View

Both mechanisms apply a row-stochastic operator to a set. In PageRank the operator G is fixed by graph structure and applied repeatedly until a global fixed point is reached; in attention the operator's rows are produced on the fly from learned query-key compatibilities and applied once to aggregate values. The attention weight vector and a PageRank transition row are the same mathematical object—a distribution over the set—so attention can be read as a single step of a learned, query-conditioned random walk, and PageRank as the converged limit of a fixed one.

$$\alpha \in \Delta^{n}, \qquad \text{row}(G_i) \in \Delta^{n}, \qquad w \geq 0,\ \sum w = 1 \tag{6}$$

Graph attention and APPNP occupy the space between these extremes: GAT learns the weights but applies them over graph neighbors like PageRank; APPNP fixes the weights as a personalized-PageRank diffusion but applies them to learned features. The personalized PageRank used by APPNP is itself a damped fixed point with a seed teleport vector e.

$$\pi_{\text{ppr}} = (1-\alpha)\,e + \alpha\,\pi_{\text{ppr}} P \ \Rightarrow\ \pi_{\text{ppr}} = (1-\alpha)(I-\alpha P)^{-1} e \tag{7}$$

Figure 1 (specification): Unifying diagram
Left: PageRank (fixed transition matrix, iterated to stationary distribution)

10.48047/jocaaa.2023.31.04.66

Center: graph attention (learned weights over neighbors, applied once/few times)
Right: self-attention (learned query-conditioned weights over all tokens, applied once)
Common label: normalize affinities to a simplex, then aggregate

Figure 1. *PageRank and attention as two ends of one weighted-aggregation spectrum.*

6. Illustrative Empirical Study: Methodology

To verify that the connection between ranking and attention is structural rather than rhetorical, we ran a small, fully reproducible study under a fixed random seed (42). It is a set of controlled computations on a synthetic directed graph and on illustrative score vectors, designed to make the shared mathematics observable rather than to benchmark systems. We computed four things: PageRank by two independent methods, the dependence of convergence on damping, the structure of attention weights, and the local-versus-global behavior of personalized PageRank.

6.1 Setup

We generated a random directed graph on twelve nodes (edge probability 0.25), converted it to a row-stochastic transition matrix with uniform handling of dangling nodes, and formed the Google matrix with damping $\alpha = 0.85$. PageRank was computed (i) by power iteration to a tolerance of 10^{-12} and (ii) as the dominant left eigenvector via eigendecomposition. We measured convergence at three damping values. For attention we applied a softmax to a peaked and a flat score vector of five elements and computed the Shannon entropy of the resulting weights. For personalized PageRank we ran a random walk with restart from a seed node at a small and a large damping value and compared the resulting rankings. Computation used NumPy and SciPy.

7. Results

7.1 PageRank's Two Computations Agree

Power iteration and eigendecomposition produced the same PageRank vector to within an L1 difference of 2.36e-13, converging in 38 iterations. This confirms operationally what the theory asserts: the stationary distribution of the random walk and the dominant eigenvector of the Google matrix are one and the same. The two methods also agreed on the top-ranked nodes.

Table 3. *PageRank computed two ways (synthetic 12-node graph, $\alpha = 0.85$).*

Quantity	Value
L1 difference (power vs. eigenvector)	2.36e-13
Power-iteration steps to 10^{-12}	38
Top-3 nodes (power iteration)	6, 0, 4
Top-3 nodes (eigenvector)	6, 0, 4

7.2 Damping Governs Convergence

Convergence slowed as the damping factor increased, exactly as the α^t error bound predicts: 18 iterations at $\alpha = 0.5$, 31 at $\alpha = 0.85$, and 37 at $\alpha = 0.95$ to reach a 10^{-10} residual. Higher damping means the walk

10.48047/jocaaa.2023.31.04.66

teleports less often, mixes more slowly, and therefore takes more power-iteration steps—the same parameter that ensures a unique solution also sets the speed of reaching it.

Table 4. *Power-iteration steps to a 10^{-10} residual vs. damping factor α.*

Damping α	0.50	0.85	0.95
Iterations to 10^{-10}	18	31	37

7.3 Attention Weights Are a Distribution Over the Set

A softmax over a peaked score vector concentrated almost all weight on one element (largest weight 0.973, entropy 0.160 nats), whereas a flat score vector spread weight nearly uniformly (entropy 1.608 nats, against the maximum of 1.609 for five elements). In both cases the weights were nonnegative and summed to one—a probability distribution over the elements, identical in type to a PageRank transition row, which we confirmed sums to one for every node. "Focused" attention is simply a low-entropy distribution; "diffuse" attention a high-entropy one.

Table 5. *Softmax attention weights for a peaked vs. flat score vector (five elements).*

Score profile	Weights	Entropy (nats)
Peaked	0.973, 0.008, 0.007, 0.007, 0.005	0.160
Flat	0.216, 0.196, 0.206, 0.186, 0.196	1.608
Uniform (max entropy)	0.20 each	1.609

7.4 Personalized PageRank Spans Local to Global

A random walk with restart from a seed node behaved as the bridge literature predicts. At low damping (α = 0.30) the distribution stayed concentrated near the seed (local), while at high damping (α = 0.95) it approached the global PageRank pattern. The two rankings were strongly but not perfectly correlated (Spearman ρ = 0.909), quantifying the tunable local-to-global trade-off that APPNP exploits to choose an effective neighborhood size [4].

Table 6. *Personalized PageRank from a seed: local (α = 0.30) vs. global (α = 0.95).*

Comparison	Value
Spearman ρ (local vs. global ranking)	0.909
Local top-3 nodes	0, 6, 4
Global top-3 nodes	0, 6, 4

7.5 Figures

Figure 2 (specification): Convergence curves
X-axis: power-iteration step
Y-axis: L1 residual (log scale)
Series: alpha = 0.50, 0.85, 0.95
Interpretation: all decay geometrically; slope steepest (fastest) for small alpha

Figure 2. *PageRank power-iteration convergence at three damping factors (from pr_results.json).*

10.48047/jocaaa.2023.31.04.66

Figure 3 (specification): Attention weight bars
X-axis: element index (1..5)
Y-axis: softmax weight (0..1)
Series: peaked scores, flat scores
Interpretation: peaked concentrates ~0.97 on one element; flat is near-uniform

Figure 3. *Softmax attention as a distribution over a set: peaked vs. flat.*

8. Discussion

The computations confirm that the link between PageRank and attention is structural. PageRank's two computations agreeing to thirteen decimal places is a reminder that the random-walk story and the eigenvector story describe one object; attention's weights forming a probability distribution shows that a single attention step is mathematically a single application of a (learned, query-conditioned) stochastic operator, the same kind of operator PageRank iterates to convergence. The entropy framing is useful: it places focused attention, diffuse attention, and a uniform random walk on one axis, the axis of how concentrated the weighting over the set is.

The differences are as instructive as the similarities. PageRank's operator is fixed by the graph and applied many times to reach a global equilibrium that is independent of any starting point; attention's operator is generated from data and applied once (per layer) to produce a context-dependent aggregation. This is why PageRank yields a single global ranking while attention yields a different weighting for every query. The bridges occupy the middle ground in exactly the way the unifying view predicts: graph attention learns the weights but, like PageRank, applies them over graph neighbors; APPNP fixes the weights as a personalized-PageRank diffusion but applies them to learned features, and its strong-but-imperfect local-to-global correlation ($\rho = 0.909$) is precisely the knob that lets it tune neighborhood size.

Seen this way, the history is less a sequence of unrelated breakthroughs than a gradual relaxation of constraints: from fixed weights and global equilibrium (PageRank), to learned weights over a fixed neighborhood (graph attention), to learned weights over an arbitrary set computed afresh per query (self-attention). Each step trades some of the closed-form, provably convergent structure of the random walk for adaptivity and expressiveness.

9. Practical Implications

- Treat attention diagnostics as distributions: entropy, concentration, and effective support are natural summaries, borrowed directly from the analysis of stochastic matrices and random walks.
- When propagation over a graph is needed, consider personalized-PageRank-based diffusion (APPNP-style) as a principled, convergent alternative to stacking many message-passing layers, which can over-smooth.
- Use Markov-chain intuition—mixing rate, teleportation, stationary behavior—to reason about iterative or recurrent attention and propagation schemes.
- Recognize that a fixed structural prior (like PageRank weights) can substitute for learned attention when data is scarce, and learned attention can generalize a fixed prior when data is plentiful.

10.48047/jocaaa.2023.31.04.66

- Report the damping or temperature that controls weight concentration; it governs both convergence (in iterated settings) and the local-global balance.

10. Limitations

This is an expository synthesis with a deliberately small, illustrative empirical component: a single synthetic graph and hand-chosen score vectors, intended to make the shared mathematics visible rather than to benchmark systems at scale. The unification is conceptual and mathematical; it does not claim that PageRank and attention are interchangeable in practice, only that they share a common structure with informative differences. The reported third-party results and architectures are summarized from their original publications. Real attention in large models also involves multiple heads, positional information, and nonlinearities that the simplified single-query view omits.

11. Future Scope

Several directions follow. First, architectures that explicitly interpolate between fixed-structure propagation and fully learned attention—choosing how much to rely on a graph prior versus learned compatibilities—could combine the data efficiency of the former with the expressiveness of the latter, building on APPNP and graph attention [3], [4]. Second, importing convergence and stability theory from Markov chains into the analysis of deep or iterative attention may yield guarantees about over-smoothing and representation collapse. Third, entropy- and spectrum-based diagnostics for attention, grounded in the stochastic-matrix view, could improve interpretability. Fourth, scaling personalized-PageRank-style propagation to very large graphs and integrating it with Transformer backbones remains an active engineering frontier.

12. Conclusion

PageRank and attention were invented for different purposes—ordering web pages and modeling language—but they compute the same kind of thing: a normalized, nonnegative weighting over a set of related elements, used to aggregate importance. PageRank fixes the weights by graph structure and iterates them to a global equilibrium; attention generates them from learned compatibilities and applies them once per query; and graph attention and personalized-PageRank propagation occupy the principled middle ground. Our reproducible study made the unity tangible: the two ways of computing PageRank agree to thirteen decimals, damping governs convergence exactly as theory predicts, and attention weights are simply a distribution over a set, the same object as a transition row. The line from PageRank to Transformers is best read not as a break but as a steady relaxation from fixed, convergent ranking toward learned, adaptive attention—one idea, progressively set free.

References

[1] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” Computer Networks and ISDN Systems, vol. 30, no. 1–7, pp. 107–117, 1998.

[2] 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), vol. 30, 2017.

10.48047/jocaaa.2023.31.04.66

[3] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in Proc. Int. Conf. Learning Representations (ICLR), 2018.

[4] J. Klicpera, A. Bojchevski, and S. Günnemann, "Predict then propagate: Graph neural networks meet personalized PageRank," in Proc. Int. Conf. Learning Representations (ICLR), 2019.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab Technical Report, 1999.

[6] A. N. Langville and C. D. Meyer, "Deeper inside PageRank," Internet Mathematics, vol. 1, no. 3, pp. 335–380, 2004.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[8] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," in Proc. Int. Conf. Learning Representations (ICLR), 2022.

[9] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2015, pp. 1412–1421.

[10] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. Int. Conf. Learning Representations (ICLR), 2017.

[11] T. H. Haveliwala, "Topic-sensitive PageRank," in Proc. 11th Int. Conf. World Wide Web (WWW), 2002, pp. 517–526.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[13] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press, 2008.

[14] R. A. Horn and C. R. Johnson, Matrix Analysis, 2nd ed. Cambridge, UK: Cambridge University Press, 2013.

[15] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," IEEE Trans. Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4–24, 2021.

[16] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," AI Open, vol. 3, pp. 111–132, 2022.