Advection Heads in an Atmosphere Foundation Model

Theodore MacMillan[1,2], **Elias Huseby**[1], **Sho Takatori**[1], **Nicholas T. Ouellette**[1]
[1]Stanford University, [2]Causal Labs
{tmacmill,huseby,stakatori,nto}@stanford.edu

Abstract

Advection describes the transport of information by a velocity field, and is a fundamental process in flow physics. We study this process in Aurora, a foundation model for atmospheric dynamics, by searching for attention heads whose attention anti-aligns with the velocity field. We discover many such "advection heads," which we find are disproportionately important for model accuracy, particularly at high wind speeds where advection dominates. Examining the attention mechanism of a low-rank advection head, we show that advection can be implemented through velocity-dependent offsets and rotations of queries and keys, causing patches to attend to their upwind neighbors. We conclude by identifying the model components responsible for these simple geometric operations.

1 Introduction

Advances in mechanistic interpretability have provided detailed, often geometric, explanations for a variety of language models behaviors including indirect object completion [1], line breaking [2], comparison circuits [3] and abbreviation creation [4]. As scientific foundation models proliferate, similar tools have begun to clarify how these models represent the physical processes they aim to predict. While this is particularly true in biological contexts [5, 6], progress has also been made in low-dimensional dynamical systems models [7, 8] and high-dimensional fluids models [9, 10].

Still, comparatively little has been shown in models capable of predicting the time evolution of partial differential equations (PDEs), and we lack any results that explicitly relate model components to physical processes. A key early result in language model interpretability was the discovery of induction heads, attention circuits capable of completing patterns by copying information from earlier in a model's context [11]. In this paper, we investigate a simple but similarly fundamental process in the context of PDEs: the transport of information by velocity fields, or advection.

Focusing on the atmosphere and ocean physics foundation model Aurora [12], we discover a significant fraction of attention heads that anti-align with the dominant wind direction; that is, downstream queries attend to upstream keys. Beyond this correlation, we find that these "advection heads" are anomalously important: removing them significantly degrades model performance, especially at the higher wind speeds where we expect advection to be physically the most important. To explore the mechanisms that may enable these kinds of attention patterns, we study a low rank advection head and describe the "QK circuit" [13] responsible for its advective behavior. Code is available at `https://anonymous.4open.science/r/advection-heads-862F`.

2 Preliminaries

2.1 Advection and Attention

The rate of change of a field $\phi(\vec{x}, t)$ transported by the velocity $\vec{v}(\vec{x}, t)$ can be decomposed as $\partial_t \phi + \vec{v} \cdot \nabla \phi = f(\phi, \vec{v}, ..)$, where $\vec{v} \cdot \nabla \phi$ describes advection and $f(\phi, \vec{v}, ...)$ is some interaction

Conference on Physics and AI at Stanford University (PAI 2026).

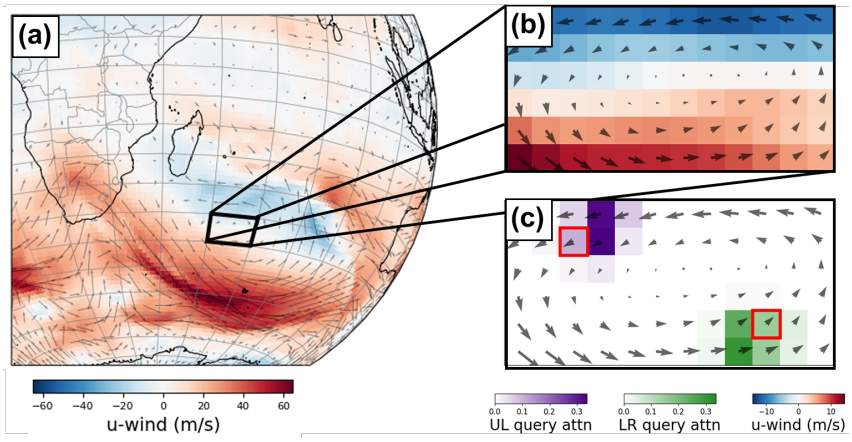

Figure 1: (a) Atmospheric velocity field at 500 hPa. Gridlines denote attention window boundaries, and color represents longitudinal velocity. (b) Zoomed in and rescaled velocity field for one window, where arrows point in the 500hPa wind direction. (c) Attention weights for two example queries, highlighted in red, with attention weights associated with an upper-left (UL) query colored in purple and a lower right (LR) query colored in green.

term. This decomposition is generic in continuum mechanics and whereas the interaction term often describes diffusion, reactions, or other types of coupling, advection can be described generally by the same simple operator.

In contrast to physical systems, attention is the *only* way in which information is mixed across tokens in a transformer. A single attention head at an arbitrary model layer has parameters $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{\text{model}} \times d_{\text{head}}}$, $\mathbf{W}_O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ and performs spatial mixing in two steps. First, each query patch i forms an attention mask over all other key patches j as $a_{ij} = \text{softmax}_j\left((x_i^T \mathbf{W}_Q^T \mathbf{W}_K x_j)/\sqrt{d_{\text{head}}}\right)$. This is the "QK" step of attention. Next, each key patch j writes to the residual stream of query patch i weighted by their attention score: $h_i = \sum_j a_{ij}(\mathbf{W}_O \mathbf{W}_V x_j)$, in a "WO" step. The QK circuit governs *where* information moves, and the WO circuit controls *what* is read and written [13].

2.2 Aurora

Aurora, the subject of our study, is a 3D Swin transformer with a U-Net architecture [14], which operates on a global grid of latitude, longitude, and atmospheric pressure levels, and is trained to predict global fields of temperature, wind, moisture, and other variables at intervals of six hours. Atmospheric foundation models have surpassed numerical solvers in accuracy out to 10 day lead times [15], and represent the state of the art in neural PDE solvers.

We direct readers to [16] for full architecture details. After an encoder perceiver transforms 3D patches of variables from a grid of longitude, latitude, and pressure onto a latent mesh, Aurora applies successive attention and MLP operations. The U-Net architecture means that the resolution of the latent mesh is additionally modified several times throughout the forward pass of the model, before a decoder perceiver eventually maps the latent mesh back onto an atmospheric grid. Critically, as the initial resolution of the latent mesh is 360×180×4 patches, attention is restricted at this stage to smaller 12×6×2 windows in the style of [17]. To allow long-range transfer of information, this attention window is shifted at every other layer.

3 Results

For each head, for each window, every query patch i attends to a distribution of key patches j. Since both query i and key j have well-defined positions on the globe—say $\vec{p}_i$ and $\vec{p}_j$—we can compute the displacement between a query and each key it attends to and add these vectors weighted by the corresponding attention weight: $\vec{\alpha}_i = \sum_j a_{ij}(\vec{p}_i - \vec{p}_j)$. This gives us a vector field of attention over every patch for every attention head.

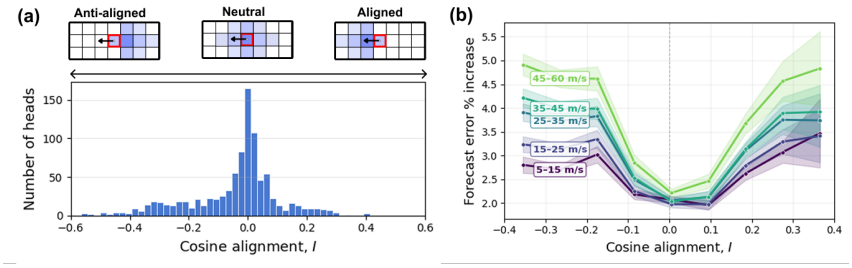

Figure 2: (a) Histogram of 928 attention heads by their alignment I. Schematics above show attention patterns for an anti-aligned, neutral, and aligned attention head relative to the velocity (black arrow). (b) Forecast error increase as a function of the alignment I, where colors from purple to green represent wind-speed from low to high. Error bars represent standard error of the mean.

We compare the attention ($\vec{\alpha}_i$) and velocity ($\vec{v}_i$) fields for each head to quantify how much the attention behaves like an advection operator. We compute alignment as the cosine similarity between these two fields, $I = \frac{1}{N}\sum_{i=1}^{N}(\vec{\alpha}_i \cdot \vec{v}_i)/|\vec{\alpha}_i||\vec{v}_i|$, where N is the number of patches, so that $I = -1$ ($I = +1$) indicates strong anti-alignment (alignment) with the velocity. We use alignment rather than magnitude difference, since the latter is dominated by large wind speeds and is less sensitive to small ones. Figure 1 shows a strongly anti-aligning head in a single window, where downstream queries attend preferentially to upstream keys.

We plot the distribution of the alignment index over all 928 attention heads in Figure 2(a). Most attention heads are uncorrelated with the velocity direction, but a significant fraction tend to look upwind (anti-align with the velocity), a behavior we expect for a head implementing advection. We also observe heads that look downwind (align with the velocity), but we do not have a physical interpretation for their behavior.

3.1 Ablations

Physically, advection begins to dominate information transfer as velocity increases. If attention heads with strong anti-alignment are implementing advection, we expect their contribution to model accuracy to grow with wind speed. To test this, we ablate each head individually and run the model for 8 different days, measuring the average increase in MSE error at 500hPa ($(\mathrm{MSE}_{500}^{\mathrm{ablated}} - \mathrm{MSE}_{500}^{\mathrm{orig}})/\mathrm{MSE}_{500}^{\mathrm{orig}} \times 100$) as a function of alignment and wind speed, displayed in Figure 2(b).

We observe that at all wind speeds, both anti-aligned and aligned heads are more important for Aurora's performance than the more typical unaligned heads, and that this gap widens at higher wind speeds. In contrast, the error associated with unaligned heads remains roughly constant. This suggests that heads with high values of alignment or anti-alignment play a key role in velocity-driven information transfer. Anti-aligned heads, which move information downwind, can be reasonably interpreted as "advection heads."

3.2 Case Study: the geometry of an advection head

In general, understanding the QK circuit of an attention head—*why* it transfers information *where* it does—can be challenging, because it may depend on many features of the data simultaneously [18]. A useful approach is to take a singular value decomposition of $\mathbf{W}_Q^T\mathbf{W}_K$ (the QK matrix), which identifies the dominant directions in the residual stream that contribute to attention scores. When the QK matrix is low rank, $R \ll d_{\mathrm{head}}$, only a few directions control most of the computation, and attention can be viewed as computing distances between queries and keys in an R-dimensional space, much simpler than the full residual stream.

Decomposing the QK matrix into its singular vectors $\mathbf{W}_Q^T\mathbf{W}_K = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, observe that the attention mask is formed as $a_{ij} = \mathrm{softmax}_j(\sum_{k=1}^{R}(\sqrt{\sigma_k}u_k^T x_i)(\sqrt{\sigma_k}v_k^T x_j))$; the attention weight depends on the distances between keys x_j in the directions of the scaled right singular vectors $\sqrt{\sigma_k}v_k$ and queries x_i in the directions of the scaled left singular vectors $\sqrt{\sigma_k}u_k$.

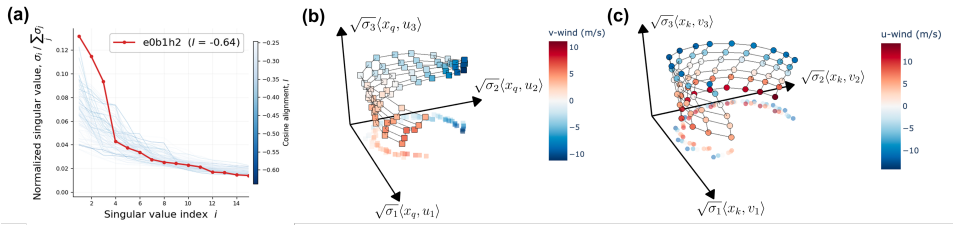

Figure 3: (a) Singular values of $W_Q^T W_K$ for the top 100 most anti-aligned attention heads. (b) Projection of the queries (keys) onto first 3 left (right) singular vectors of the QK matrix of e0b1h2 for each patch in a sample window.

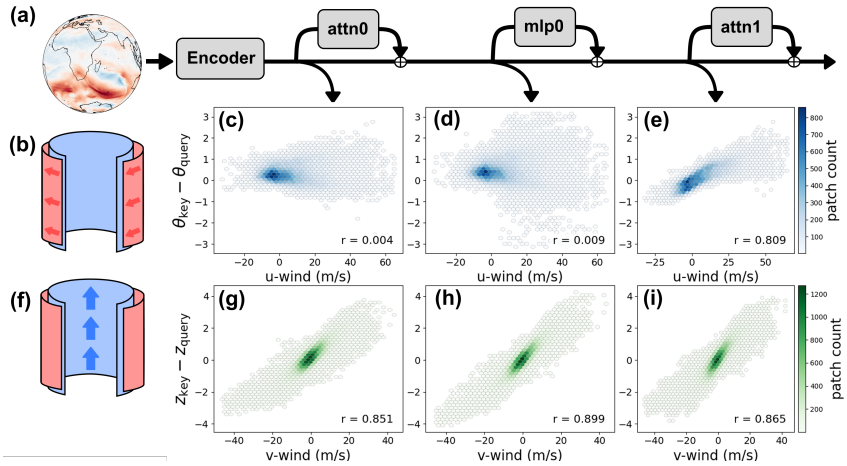

Figure 4: (a) Partial schematic of the Aurora pipeline where an atmospheric state is first encoded and then processed by interwoven attention and MLP blocks. (b) Schematic depicting queries (red) rotating with respect to keys (blue) in the case of a constant longitudinal velocity. (c-e) Distribution of polar offsets between keys and queries for each patch after each addition to the residual stream. (f) Schematic depicting keys vertically offset from queries due in the case of a constant meridional velocity. (g-i) Distribution of vertical offsets between keys and queries for each patch after each addition to the residual stream.

Figure 3(a) shows the singular values of the top 100 most advection aligned heads. We find head e0b1h2 (encoder layer 0, block 1, head 2) has substantially lower effective rank than most advection heads, meaning we can likely interpret its QK circuit. Figure 3(b) and 3(c) show the keys and queries projected into these SV-aligned directions of e0b1h2 for a single representative window (the same displayed in Figure 1). For both keys and queries, the representations resemble truncated cylinders, where longitude is encoded angularly and latitude is encoded vertically. If this were the only step, each query would attend primarily to itself and its neighbors. However, we also find that the keys are vertically offset by meridional velocity and that the queries are rotated by zonal velocity. In this way, queries are able to attend to their upstream neighbors in both relevant directions.

To understand where in the model these offsets emerge, we can look at the residual stream projected onto these SV-aligned directions at every stage prior to e0b1h2. Figure 4 shows the angular and vertical offset between every patch between -60° and 60° latitude on its key side and query side as functions of zonal and meridional wind, respectively. We find that the vertical offset is implemented in the initial encoding module, but that MLP0 is responsible for the twist that allows the zonal attention patterns to form. Although this description offers only a limited window into Aurora's remarkable performance, that such a description is possible is encouraging for future mechanistic analysis of PDE foundation models.

References

[1] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. 11 2022.

[2] Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025.

[3] Jack Merullo, Connor Watts, Max Loeffler, Liv Gorton, Elana Simon, Tom McGrath, and Owen Lewis. Replicating circuit tracing for a simple known mechanism, June 2025.

[4] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

[5] Elana Simon and James Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 10 2025.

[6] Kevin Lu, Jannik Brinkmann, Stefan Huber, Aaron Mueller, Yonatan Belinkov, David Bau, and Chris Wendler. Mechanisms of AI Protein Folding in ESMFold. 2 2026.

[7] Anthony Bao, Venkata Hasith Vattikuti, Jeffrey Lai, and William Gilpin. Universal redundancies in time series foundation models. *arXiv preprint arXiv:2602.01605*, 2026.

[8] Anthony Bao, Jeffrey Lai, and William Gilpin. Transformers for dynamical systems learn transfer operators in-context. *arXiv preprint arXiv:2602.18679*, 2026.

[9] Rio Alexa Fear, Payel Mukhopadhyay, Michael McCabe, Alberto Bietti, and Miles Cranmer. Physics steering: Causal control of cross-domain concepts in a physics foundation model. *arXiv preprint arXiv:2511.20798*, 2025.

[10] Theodore MacMillan and Nicholas T Ouellette. Towards mechanistic understanding in a data-driven weather model: internal activations reveal interpretable physical features. *arXiv preprint arXiv:2512.24440*, 2025.

[11] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[12] Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, 5 2025.

[13] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[14] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[15] Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R. Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Skillful joint probabilistic weather forecasting from marginals. 6 2025.

[16] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[18] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy$_m odel/index.html$.