



What Does GPT-2 Look At?

RESEARCH REPORT

Chen En Lo*
Cerulean Labs

Ben*
Cerulean Labs

Abstract

Understanding the attention patterns of pretrained language models can help characterize their internal behavior. Inspired by prior analyses of BERT, we examine relative-position preferences, attention to token categories, attention entropy, and head similarity in GPT-2 Small. Across 1,000 Wikipedia segments, we identify five heads that assign more than half their attention to the current token and two that strongly prefer the preceding token. Many heads in intermediate and deeper layers attend heavily to an initial end-of-text (EOT) token, whereas non-boundary tokens assign almost no attention to natural paragraph boundaries. Entropy is relatively high in early layers, and trailing-EOT queries exhibit a distinct layer-dependent trend. Finally, a two-dimensional embedding reveals clusters associated with different attention patterns, with earlier-layer heads generally more dispersed than later-layer heads.

1 Introduction

The Transformer architecture Vaswani et al. [2017] underlies many successful large language models. Its attention mechanism enables tokens to incorporate information from other positions in a sequence. Analyzing attention patterns in a pretrained model can help characterize its learned behavior, although attention weights alone do not establish the functional role of a head.

Clark et al. [2019] proposed several methods for analyzing attention heads and applied them to BERT. Inspired by this work, we apply a subset of these methods to GPT-2 and compare the observed patterns with those reported for BERT.

We first measure how much attention each head assigns to the current token and the preceding token. We then examine attention to the initial EOT token, paragraph boundaries, and punctuation. Next, we use the entropy of attention distributions to characterize how broadly heads attend. Finally, we visualize pairwise similarities between heads in a two-dimensional space.

Our analysis addresses three questions: Do attention heads favor particular relative positions? Do they preferentially attend to specific token types? And do heads within or across layers exhibit similar attention patterns?

2 Surface-Level Patterns

We analyze surface-level patterns in GPT-2’s attention heads. Figure 1 illustrates several examples.

Setup. We extract attention maps from GPT-2 Small over 1,000 randomly sampled Wikipedia segments, each containing at most 1,024 tokens. The relative-position heat maps in Figure 2 aggregate 381,458 token positions across these segments. Most heads assign relatively little attention to the current or immediately preceding token, although a few show strong local preferences.

*Equal contribution.

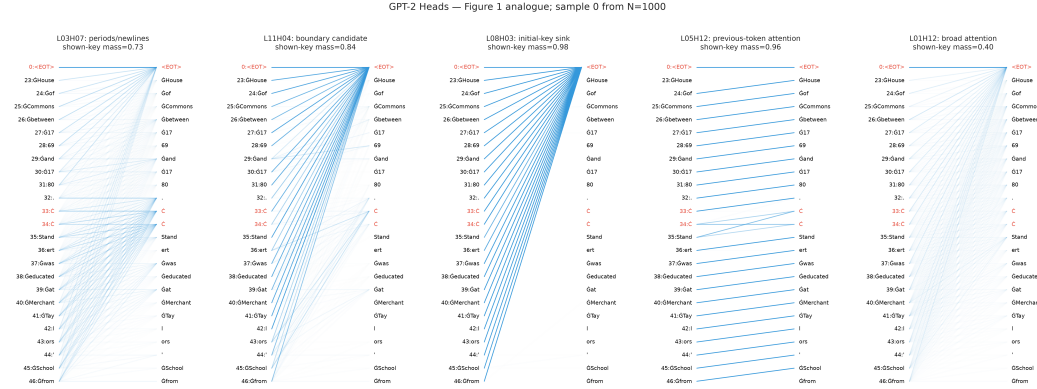


Figure 1: Illustrative attention-head patterns for a single Wikipedia segment, including attention to punctuation/newlines, the initial EOT token, the preceding token, and broadly distributed attention. Darker lines indicate higher weights; these examples should not be confused with averages across the full dataset. Causal masking prevents attention to future positions.

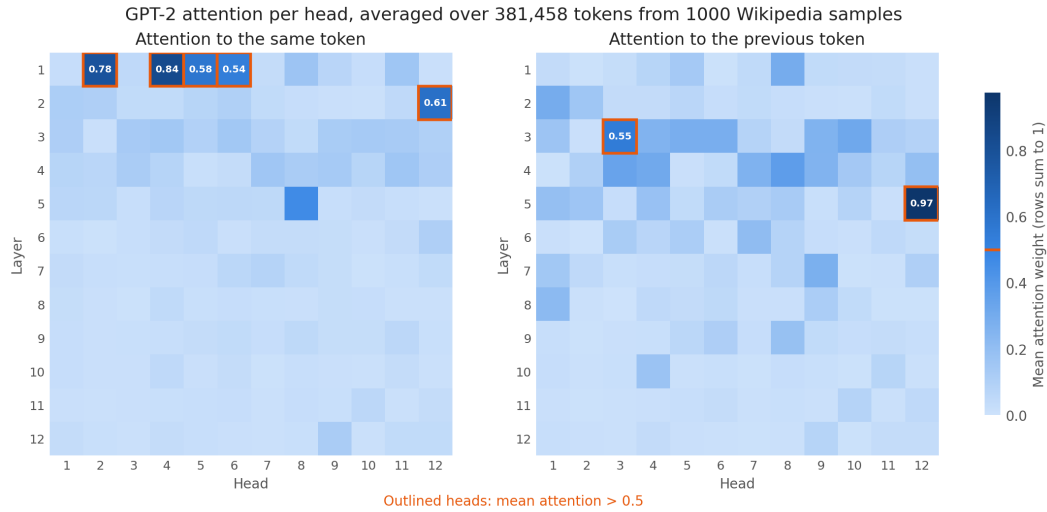


Figure 2: Mean attention assigned to the current token (left) and immediately preceding token (right), aggregated over 381,458 token positions from 1,000 Wikipedia segments. Orange outlines mark heads whose mean attention to the specified position exceeds 0.5. Layers and heads are numbered from 1.

2.1 Relative Position

We compute the average attention assigned by each head to the current token and the immediately preceding token. Most heads assign relatively little attention to either position, but a small number show pronounced local preferences. As highlighted in Figure 2, five heads assign more than 50% of their average attention to the current token: L1H2, L1H4, L1H5, L1H6, and L2H12. In contrast, two heads preferentially attend to the preceding token: L3H3 (0.55) and L5H12 (0.97). Here $LxHy$ denotes head y in layer x , using one-based indexing. Attention to a subsequent token is zero by construction because GPT-2 uses a causal attention mask.

2.2 Where Does GPT-2 Attend?

Clark et al. [2019] suggested that attention to BERT’s [SEP] token may serve a no-op function. Attention weights alone do not establish whether an analogous mechanism exists in GPT-2. We therefore focus on describing the attention patterns observed in our experiments.

We examine the initial end-of-text (EOT) token in the analyzed sequences. Under causal masking, a trailing EOT token cannot receive attention from preceding positions, whereas an initial EOT token is visible to subsequent positions. We also examine attention to natural paragraph boundaries (`\n\n`) and to punctuation (periods and commas).

The upper panel of Figure 3 shows that many heads allocate substantial attention to the initial EOT token, far more than to paragraph boundaries or punctuation. The layer-wise average rises through the earlier layers, remains high across much of the middle and later network, and decreases somewhat in the final layer; the trend is not monotonic. For example, head L8H3 assigns approximately 0.93 of its mean attention to the initial EOT token.

Although non-boundary tokens assign almost no attention to paragraph boundaries, this does not imply that boundaries are ignored by every query. The middle panel of Figure 3 separates non-boundary-to-boundary attention from boundary-to-boundary attention (including self-attention). The former remains near zero across layers. The latter is markedly higher in some early-layer heads, exceeding 0.8 in individual cases, before generally decreasing in later layers. This measurement includes attention between boundary positions as well as self-attention; it should not be described as self-attention alone.

2.3 Focused vs. Broad Attention

To quantify the concentration of attention, we compute the entropy of each head’s attention distribution (Figure 4). In the upper panel, mean entropy is high in the first two layers, drops through the intermediate layers, and rises again toward layer 12. Several early-layer heads have particularly high entropy, indicating that their attention is spread over many available key positions.

The lower panel of Figure 4 instead uses only the trailing EOT token as the query position. Its layer-wise mean entropy is relatively high in the first few layers but substantially lower through layers 4–11 than in the all-query panel; it rises again at layer 12. The two panels also display different causal-uniform reference levels. Thus, while the plotted raw entropies reveal distinct trends, the absolute magnitudes should not be interpreted as directly comparable measures of attention concentration without accounting for the number of available key positions.

3 Clustering Attention Heads

Are attention heads in the same layer similar to one another? Can heads be grouped according to their behavior? To investigate these questions, we compute pairwise dissimilarities between attention heads. For two heads H_i and H_j , we define the aggregate dissimilarity as

$$D(H_i, H_j) = \sum_{t \in \mathcal{T}} \text{JS}(H_i(t), H_j(t)),$$

where $\mathcal{T}$ is the set of query-token instances in the analyzed data, $H_i(t)$ denotes the attention distribution produced by head i at query position t , and JS is the Jensen–Shannon divergence. We apply multidimensional scaling Kruskal [1964] to embed the heads in two dimensions, approximately preserving their pairwise dissimilarities.

Figure 5 shows groups corresponding to the behaviors described in Section 2. The two preceding-token specialists appear at the bottom of the left panel, the five current-token specialists lie farther to the left, and many broadly attending heads appear in the upper-left region. These observations are consistent with the head counts in Figure 2. For the plotted behavior labels, a head is classified as broadly attending when its entropy exceeds the mean by more than one standard deviation.

Coloring the embedding by layer reveals considerable mixing across layers rather than a clean layer-wise partition. Heads from early layers are relatively dispersed, whereas many later-layer heads occupy a more concentrated region on the right of the plot, although overlap remains. These trends suggest greater variation among early-layer attention patterns, but an MDS embedding alone cannot establish a shared computational function for deeper-layer heads.

GPT-2 Heads — Figure 2 analogue (N=1000)

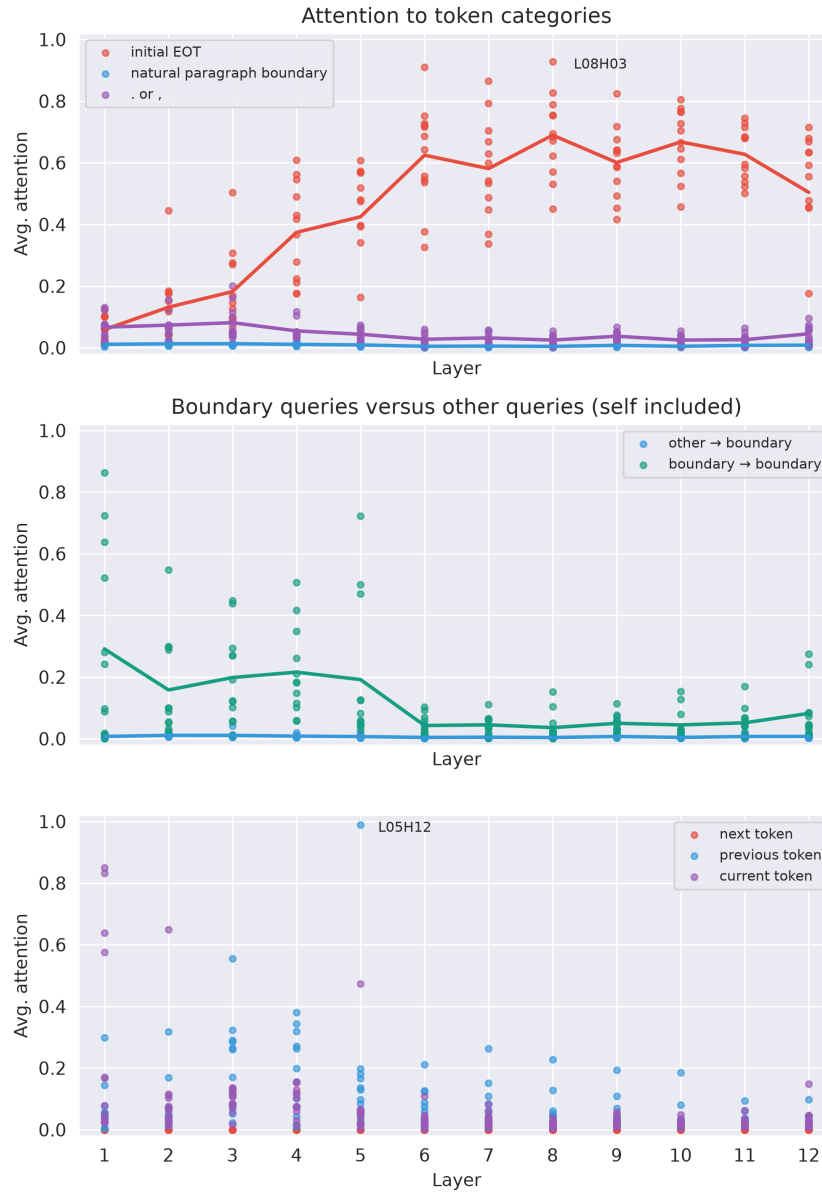


Figure 3: Mean attention by head and layer. Top: attention to the initial EOT token, natural paragraph boundaries, and periods/commas; initial-EOT attention is prominent across middle and later layers. Middle: non-boundary-to-boundary attention is nearly zero, while boundary-to-boundary attention (self-attention included) can be substantial for some early-layer heads. Bottom: attention to the next, preceding, and current tokens; next-token attention is zero under causal masking, and only a few heads show strong current- or previous-token preferences.

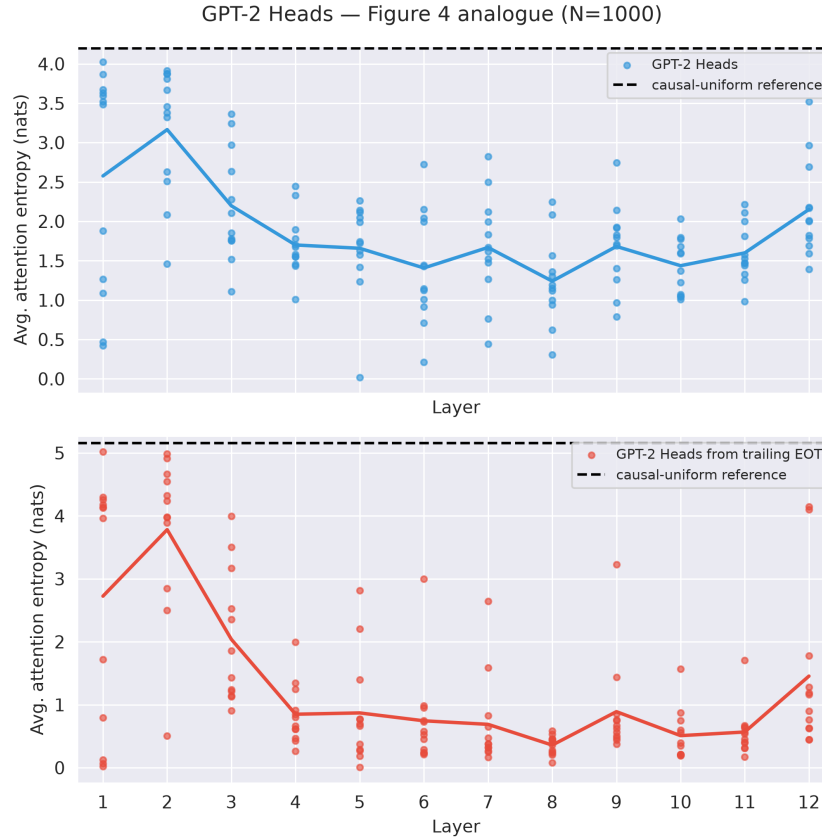


Figure 4: Mean attention entropy (nats) by head and layer. Top: entropy averaged over the analyzed query positions, with high values in early layers and a smaller rise toward the last layer. Bottom: entropy when only the trailing EOT is queried, which is lower across much of layers 4–11 and rises at layer 12. Dashed lines denote the causal-uniform reference for each panel; their different levels caution against direct comparison of unnormalized entropy values.

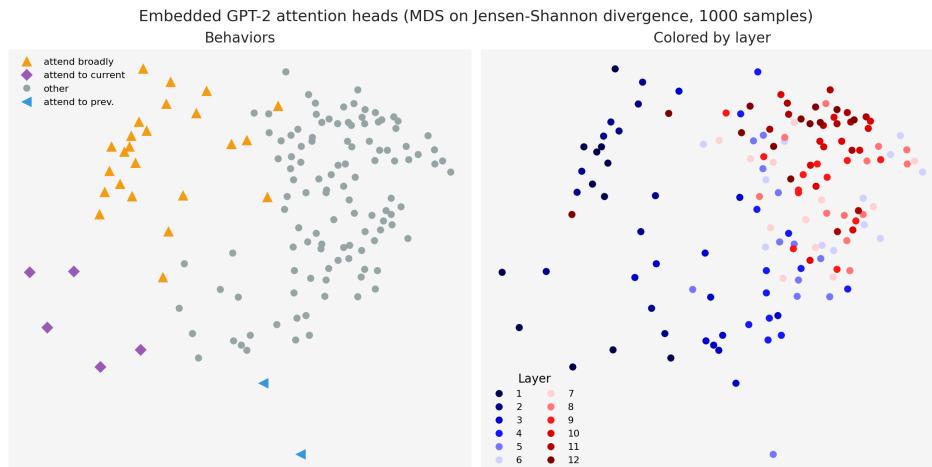


Figure 5: Two-dimensional multidimensional-scaling embedding of GPT-2 heads using aggregate Jensen–Shannon dissimilarities. Left: behavior-based labels distinguish broadly attending heads, five current-token specialists, and two preceding-token specialists. Right: colors indicate layers; early heads are more dispersed and many deeper heads are concentrated toward the right, but the layers are not cleanly separated.

4 Discussion and Conclusion

We compare our GPT-2 observations with attention patterns reported for BERT by Clark et al. [2019], distinguishing architectural constraints from empirical differences whose causes remain uncertain.

4.1 Relative Position

GPT-2’s causal attention mask excludes future positions, so attention to the next token is zero by construction. Unlike BERT’s bidirectional attention, GPT-2 can exhibit strong local preferences only for the current and earlier positions (Figure 2). This is an architectural constraint, not a learned preference inferred from our experiments.

We find five current-token specialists in layers 1–2 and two preceding-token specialists in layers 3 and 5 (Figure 2). The latter observation is consistent with the two preceding-token specialists displayed at the bottom of Figure 5. In contrast to GPT-2, BERT exhibits both preceding- and next-token specialists because it uses bidirectional attention [Clark et al., 2019].

4.2 Where Does GPT-2 Attend?

In GPT-2, attention to the initial EOT token becomes prominent in middle and later layers, though it drops somewhat in the final layer (Figure 3). Paragraph-boundary tokens receive almost no attention from non-boundary queries, but some boundary queries assign substantial attention to boundary positions, especially in early layers. Therefore, describing all paragraph-boundary attention as negligible would be misleading.

For BERT, Clark et al. [2019] reported a different layer-dependent pattern: some early-layer heads attend to [CLS], middle-layer heads often attend strongly to [SEP], and later-layer heads place more attention on periods and commas. In GPT-2, periods and commas generally receive low attention across layers, rather than exhibiting a comparable increase near the final layers. However, these comparisons are descriptive: BERT’s dedicated separator token and GPT-2’s natural paragraph boundary are not equivalent input markers.

4.3 Focused vs. Broad Attention

The all-query GPT-2 entropy curve shares a qualitative trend with the BERT results of Clark et al. [2019]: relatively high entropy in early layers, a reduction through the middle, and a rise toward the end. However, GPT-2’s trailing-EOT-only query displays a different profile, with relatively low raw entropy in layers 4–11 followed by an increase in layer 12 (Figure 4). Furthermore, the different causal-uniform reference levels in the two GPT-2 panels and the different sequence lengths used for BERT and GPT-2 preclude a direct comparison of raw entropy magnitudes.

4.4 Clustering Attention Heads

In GPT-2, heads do not form clearly separated clusters by layer, especially in early layers. However, the embeddings of later-layer heads tend to be more concentrated (Figure 5). By comparison, Clark et al. [2019] reported that BERT heads within the same layer often behave similarly. In both models, heads with certain attention patterns can also form clusters across layers.

4.5 Effect of Causal Architecture

One architectural difference between GPT-2 and BERT is GPT-2’s causal mask, which prevents attention to future positions. This directly explains the absence of next-token attention in GPT-2. Although we observe other differences in attention to special tokens, punctuation, and layer-wise head similarity, our experiments do not isolate the effects of causal masking from other differences in model architecture, tokenization, or input preparation. We therefore do not attribute those observations specifically to causal masking.

4.6 Conclusion

Across the analyzed Wikipedia segments, five GPT-2 heads strongly attend to the current token and two to the preceding token, while causal masking excludes attention to future positions. Initial-EOT attention becomes strong in the intermediate and later layers, whereas paragraph boundaries are largely ignored by non-boundary queries but not necessarily by boundary queries themselves. Attention entropy depends on query selection, and the head embedding reveals behavior-based groups rather than clearly separated layer-wise clusters. These findings characterize observed attention patterns; establishing the functional roles of these heads requires further controlled experiments.

5 Limitations

Our analysis has several limitations. First, describing attention patterns does not establish the computational role of individual heads. Patterns such as local attention, strong attention to an initial EOT token, or high-entropy distributions may reflect multiple mechanisms. In particular, the initial EOT token is confounded with its position at the beginning of the sequence. Controlled interventions would be needed to distinguish effects of token identity from positional effects and to test whether these patterns causally influence model behavior.

Second, our experiments use only Wikipedia segments and GPT-2 Small. The results may not generalize to other text domains, such as dialogue, code, or news, or to different model sizes. A controlled GPT-2–BERT comparison would require accounting for attention masks, tokenization, input markers, and sequence lengths. Even within our GPT-2 results, differences in the number of valid attention keys affect raw entropy, motivating length-aware or normalized comparisons. Finally, an MDS visualization preserves pairwise dissimilarities only approximately and is not a direct measure of the causal function of individual attention heads.

References

- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert’s attention, 2019. URL <https://arxiv.org/abs/1906.04341>.
- Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964. doi: 10.1007/BF02289565.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.