What are the characteristics of attention?

Alignment with syntax
How does attention correlate to syntactic properties?

Distance
Does attention capture short-range or long-range relationships?

Entropy
Is attention dispersed broadly over many tokens or focused on a few?

Variability
How does attention vary over inputs?

Exemplars
Which sentences/tokens most strongly induce attention in a particular head?

Experiment
Sample sentences ➔ Run GPT-2 on each sentence ➔ Extract attention ➔ Compute assorted metrics ➔ Aggregate over corpus

Key Results

Attention heads target very specific lexical patterns
For each attention head, we identified the sentences (exemplars) that most strongly induced attention in that head. Below we show the top exemplar for each of 3 heads, along with the arcs with maximum attention. Other exemplars for each head followed similar patterns.

Layer 10 / Head 10: The Australian search and rescue service is provided by Aus S AR, which is part of the Australian Maritime Safety Authority (AM S A ).
Connections: acronym to associated phrase (likely for predicting next acronym piece)

Layer 11 / Head 2: After the two prototypes were completed, production began in May at 1a, Georgia, where over 2 130 C-130s have been built...
Connections: acronym to preceding place name (likely for predicting following place name)

Layer 11 / Head 10: ... some scale in World War I, the prospects of Anglo-American assistance in another war with Germany appeared to be doubtful ...
Connections: end of noun phrase to head word (likely for predicting following verb)

Model View
Visualizes attention across all of the model’s layers and heads for a particular input.

Attention-Head View
Visualizes attention in one or more heads for a given layer.

Neuron View
Shows how attention is computed from query and key vectors.

Mean Attention Distance by Layer / Head

Mean Attention Distance by Layer / Head

Mean Attention Distance (Non Tokens)

Dependency Alignment by Layer / Head

Dependency Alignment by Layer / Head

Attention aligns with syntactic dependencies most strongly in the middle layers of the model.

The deepest layers capture the longest-range relationships.

Attention heads specialize in different parts of speech at different layer depths.

Heads attend more to lower-level parts of speech in lower layers

Determiner

Heads attend more to higher-level parts of speech in deeper layers

Proper Noun

Each heatmap shows the proportion of total attention focused on the respective part of speech, broken out by layer (vertical axis) and head (horizontal axis).

Attention targets key vectors.
For each attention head, we identified the sentences (exemplars) that most strongly induced attention in that head. Below we show the top exemplar for each of 3 heads, along with the arcs with maximum attention. Other exemplars for each head followed similar patterns.

Layer 10 / Head 10: The Australian search and rescue service is provided by Aus S AR, which is part of the Australian Maritime Safety Authority (AM S A ).
Connections: acronym to associated phrase (likely for predicting next acronym piece)

Layer 11 / Head 2: After the two prototypes were completed, production began in May at 1a, Georgia, where over 2 130 C-130s have been built...
Connections: acronym to preceding place name (likely for predicting following place name)

Layer 11 / Head 10: ... some scale in World War I, the prospects of Anglo-American assistance in another war with Germany appeared to be doubtful ...
Connections: end of noun phrase to head word (likely for predicting following verb)

The deepest layers capture the longest-range relationships.

Mean attention distance is the mean distance (in number of tokens) between all pairs of tokens, weighted by the attention between the tokens.

Entropy

Entropy

Entropy

Mean Entropy

Mean Attention Distance by Layer / Head

Mean Attention Distance by Layer / Head

Mean Attention Distance (Non Tokens)

Dependency Alignment by Layer / Head

Dependency Alignment by Layer / Head

Attention aligns with syntactic dependencies most strongly in the middle layers of the model.

The deepest layers capture the longest-range relationships.

Mean attention distance is the mean distance (in number of tokens) between all pairs of tokens, weighted by the attention between the tokens.

Entropy

Entropy

Entropy

Mean Entropy

Mean Attention Distance by Layer / Head

Mean Attention Distance by Layer / Head

Mean Attention Distance (Non Tokens)

Dependency Alignment by Layer / Head

Dependency Alignment by Layer / Head

Attention aligns with syntactic dependencies most strongly in the middle layers of the model.

The deepest layers capture the longest-range relationships.

Mean attention distance is the mean distance (in number of tokens) between all pairs of tokens, weighted by the attention between the tokens.

Entropy

Entropy

Entropy

Mean Entropy

Mean Attention Distance by Layer / Head

Mean Attention Distance by Layer / Head

Mean Attention Distance (Non Tokens)

Dependency Alignment by Layer / Head

Dependency Alignment by Layer / Head

Attention aligns with syntactic dependencies most strongly in the middle layers of the model.

The deepest layers capture the longest-range relationships.

Mean attention distance is the mean distance (in number of tokens) between all pairs of tokens, weighted by the attention between the tokens.

Entropy

Entropy

Entropy

Mean Entropy

Mean Attention Distance by Layer / Head

Mean Attention Distance by Layer / Head

Mean Attention Distance (Non Tokens)

Dependency Alignment by Layer / Head

Dependency Alignment by Layer / Head

Attention aligns with syntactic dependencies most strongly in the middle layers of the model.

The deepest layers capture the longest-range relationships.

Mean attention distance is the mean distance (in number of tokens) between all pairs of tokens, weighted by the attention between the tokens.