Task Optimization Leads to Human-like Top-down and Bottom-up Attention during Reading Comprehension

Jiajie Zou
College of Biomedical Engineering and Instrument Sciences, Zhejiang University
Nai Ding (ding_nai@zju.edu.cn)
College of Biomedical Engineering and Instrument Sciences, Zhejiang University

Article

Keywords: reading comprehension, information processing, human attention, deep neural network

Posted Date: November 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1034025/v1

Task Optimization Leads to Human-like Top-down and Bottom-up Attention during Reading Comprehension

Jiajie Zou[1], Nai Ding[1,2*],

[1] Key Laboratory for Biomedical Engineering of Ministry of Education,
College of Biomedical Engineering and Instrument Sciences,
Zhejiang University, Hangzhou 310027, China

[2] Zhejiang lab, Hangzhou 311121, China

***Corresponding author:**

Nai Ding,

Email: ding_nai@zju.edu.cn

Zhejiang lab, Hangzhou 311121, China

Abstract

Attention is a complex system involving multiple interactive components that jointly regulate information processing in the brain. It has been hypothesized that the computational goal of attention is to optimally integrate information under task demands, and evidence has been provided in relatively simple learning and decision making tasks. It remains unclear, however, whether this hypothesis can explain attention distribution in more complex real-world tasks that engage multiple attention systems. Here, taking advantage of the development of attention mechanisms in deep neural network (DNN) models, we investigate whether human attention during real-world reading comprehension tasks can be explained as a consequence of task optimization. In a goal-directed reading task, participants read a passage to answer a question. Eye tracking results show that the attention on each word, quantified by the fixation time, is modulated by both the top-down reading goal and lower-level visual layout and textual features. When trained to perform the same goal-directed reading task, DNN models yield human-level performance and naturally evolve human-like attention distribution, with deep layers tuned to the reading goal and shallow layers tuned to textual features. Further experiments suggest that different training tasks separately contribute to goal-directed and text-based attention. In summary, the results strongly suggest that human attention can be interpreted as a consequence of task optimization during real-world reading tasks.

Introduction

Attention is a key mechanism to regulate information processing in the brain and exerts profound influences on perception and cognition[1,2]. Instead of a single module, the attention system can be functionally divided into multiple components[3-5]. For example, goal-directed top-down attention is mainly guided by the task, while stimulus-driven bottom-up attention is mainly determined by the physical features of stimuli[6]. In terms of the neural implementation, it has been suggested that the control of top-down and bottom-up attention engages separable cortical pathways[7,8] and these attention control pathways are widly connected to other cortical and subcortical areas to effectively control neural processing across sensory modalities at multiple levels along the processing hierarchy[9-11].

Since attention has multiple function components and engages broad cortical areas, it remains unresolved whether attention plays a unified computational role in information processing. Traditionally, attention is viewed as a mechanism to allocate limited central processing resources[12,13]. More recent studies, however, propose that attention is a mechanism to optimize learning and task performance, even in conditions where the processing resource is not clearly constrained[14-17], and this optimization view has been used to explain the attention distribution in a range of learning and decision making tasks[18-20]. Little empirical evidence has been provided, however, whether task optimization can explain attention distrubition in more

ecologically valid complex tasks that engage multiple forms of attention, since it is generally challenging to solve optimization problems for complex tasks.

Here, taking advantage of recent development of attention mechanisms in DNN[21,22], we investigate whether task optimization is sufficient to explain human attention during a real-world challenging reading task, i.e., answering high-school level reading comprehension questions[23]. DNN are powerful tools to solve complex optimization problems and, using DNN previous studies have shown that task optimization can indeed explain, for example, the response properties of individual neurons[24-26]. We choose the reading comprehension task for 3 reasons. First, the task strongly engages attention, since finding the answer to a question in a long passage poses a challenge for information selection. Second, the task is of high ecological validity: Information searching is one of the most common reading activities in the modern time[27] and people often receive practice for this task in school. Third and importantly, DNNs equipped with the self-attention mechanism have reached human-level performance on this task.

The purpose of our study is to investigate whether DNNs optimized for task performance can generate attention distribution that is quantitatively similar to the attention distribution of humans who perform the same task. We quantified human attention through eye tracking and quantified DNN attention using the attention weight on each word. In the following, we first report how human attention is

modulated by stimulus features and the top-down reading goal, and then analyze whether DNN evolves bottom-up and top-down attention like humans. Lastly, we analyze how the training tasks contribute to the distribution of DNN attention and and its similarity to human attention.

Results

Human Attention Distribution and Influence Factors

In Study 1, the participants (N = 25 for each question) first read a question and then read a passage based on which the question should be answered (Fig. 1A). After reading the passage, the participants read 4 options related to the question and had to choose which option was the most suitable answer. Eight hundred question/passage pairs were presented, and the questions fell into two broad categories, i.e., local and global questions (see *Materials and Methods* for details). Local questions require attention to details while global questions concern the general understanding of a passage. The participants correctly answered 77.94% questions on average (Fig. 1E, 77.49% and 78.77% for local and global questions, respectively).

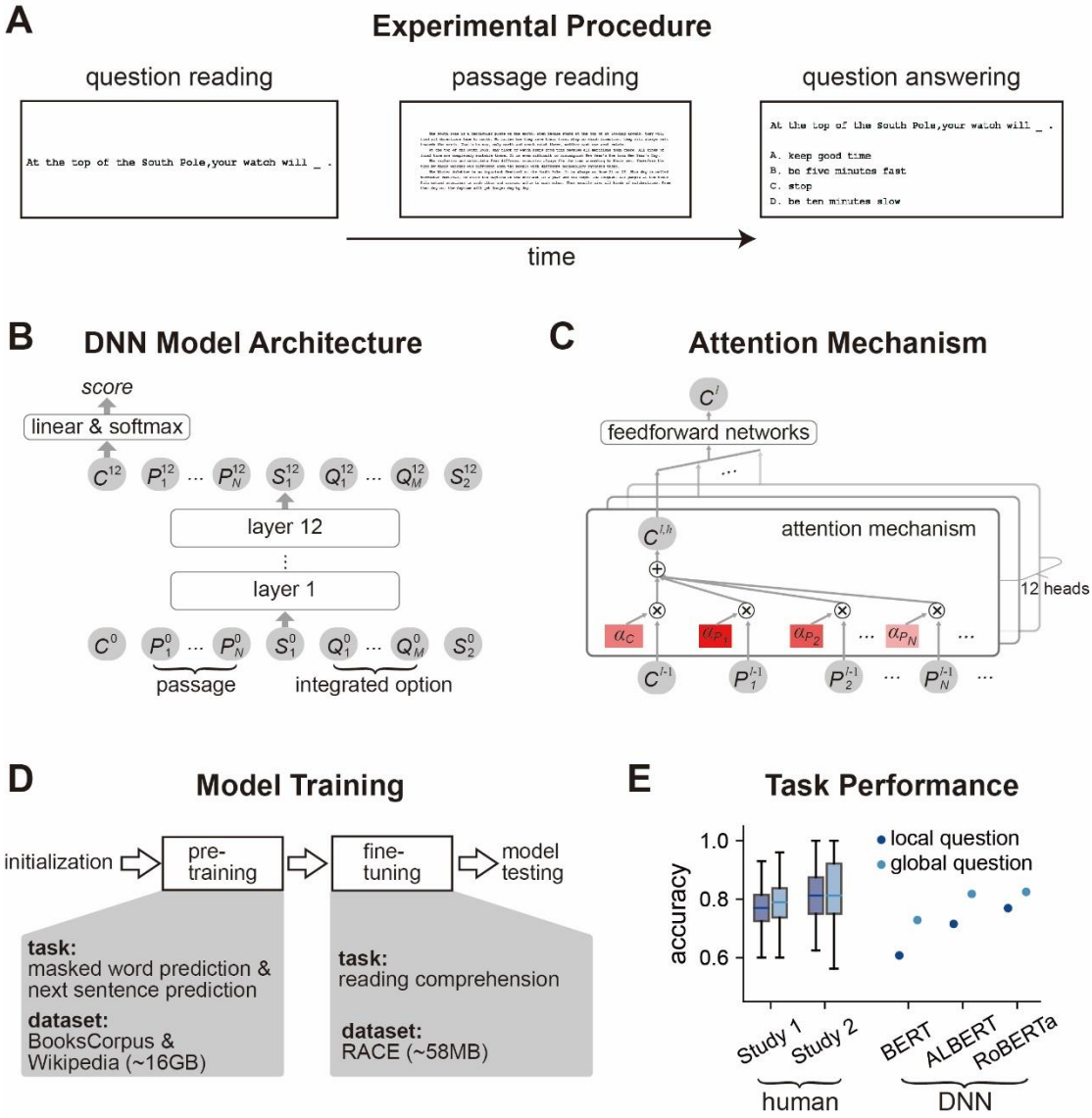


Fig. 1. Experimental procedure and DNN model. (**A**)The experimental procedure in Study 1. In each trial, participants read a question first, and then read the corresponding passage, and finally proceeded to read and answer the question which is coupled with 4 options. Here, the question and options are shown in a larger font size for illustration purposes. (**B**) Architecture of the DNN models used for the reading comprehension task. The input to the models consists of all words in the passage and an integrated option, and also 3 special tokens, i.e., CLS, SEP_1, and SEP_2 (denoted as C, S_1, and S_2), separating the passage and the integrated option. The CLS token is the decision variable and its final representation is used to calculate a score that reflects how likely the option is the correct answer. The DNN model has 12 layers, and each layer has 12 parallel self-attention modules, i.e., attention heads. (**C**) Illustration of the DNN attention mechanism in a layer. In the models, each word/token is represented by a vector, and information is integrated across

words/tokens only in the self-attention module. The vectorial representation of the CLS token is a weighted sum of the vectorial representations of all words and tokens, and the attention weight for each word in the passage, i.e., α_{Pn}, is the DNN attention analyzed in this study. Output of the self-attention model, i.e., $C^{l,h}$, is further processed by feedfoward networks and other operations that do not engage information integration across words. (**D**) DNN models are trained in two steps. The pre-training process aims to learn general statistical regularities in a language based on large corpora, while the fine-tuning process trains models to perform the reading comprehension task. (**E**) Performance of humans and DNN models on the reading comprehension task.

While the participants read the passage, their eye gaze was monitored using an eye tracker, and their attention to each word was quantified by the total fixation time on the word. The results showed that longer words were fixated for longer time (Fig. S1), consistent with previous studies[28]. Nevertheless, the fixation time on a word was expected to be positively related with the area the word occupied even when attention was uniformly distributed across the visual field. Therefore, here we further extracted the *attention density* by dividing the total fixation time on a word by the area the word occupied, and used this measure in subsequent analyses. The attention density clearly deviated from a uniform distribution (see Fig. 2 for examples). To probe into the factors modulating human attention distribution, we quantified how the human attention distribution was influenced by multiple sets of features in the following.

We first analyzed whether textual features, e.g., word length, word frequency, and a word's position in a sentence, could predict human attention distribution using linear

143 regression. The prediction accuracy, i.e., the correlation coefficient between the
144 predicted and actual attention density, was significantly above chance (P = 0.002,
145 permutation test, FDR corrected). Furthermore, the prediction accuracy was
146 significantly higher for global questions than for local questions ($P = 1.4 \times 10^{-4}$,
147 bootstrap, FDR corrected) (Fig. 3A, the left plot). We then used the same regression
148 analysis to analyze whether the visual layout of a passage could also affect attention
149 distribution. Here, layout features referred to features induced by line changes (see
150 *Materials and Methods* for details), which could be processed without word
151 recognition. The prediction accuracy for layout features was also statistically
152 significant (P = 0.002, permutation test, FDR corrected).
153
154 Textual features and layout features characterized properties of the stimulus that were
155 invariant across tasks. In the following, we investigated whether the task, i.e., to
156 answer a specific question, also modulated human attention distribution. To
157 characterize the top-down influence of task, we acquired annotations indicating each
158 word's contribution to question answering, i.e., task relevance (see *Materials and*
159 *Methods*). As shown in the left plot of Fig. 3A, we found that task relevance could
160 indeed significantly predict human attention distribution (P = 0.002, permutation test,
161 FDR corrected). Since task relevance was not a well-established modulator of reading
162 attention, we further analyzed whether the task relevance effect could be explained by
163 the well-established textual and layout effects. In this analysis, we first regressed out
164 the influence of textual and layout features from the human attention distribution, and

found that the residual attention distribution could still be predicted by task relevance ($P = 0.003$, permutation test, FDR corrected) (Fig. 3A, middle plot). These results showed that the top-down reading goal, quantified by task relevance, could modulate human attention, on top of lower-level stimulus features, i.e., textual and layout features.

The linear regression analyses revealed that textual features, layout features, and task relevance all modulated human attention (see Fig. 2 for examples). The prediction accuracy for different features ranged between 0.2 and 0.6, comparable to the prediction accuracy of visual saliency models when predicting human attention to images[29,30]. Further analyses also revealed how these features modulated human attention. For example, we found that participants generally attended more to the beginning of a passage (Fig. 3C). Furthermore, this effect was stronger for global questions, which potentially explained why stimulus features could better predict the attention distribution for global questions. Additionally, it was also found that participants attended more to words that are more relevant to the question answering task (Fig. 3D).

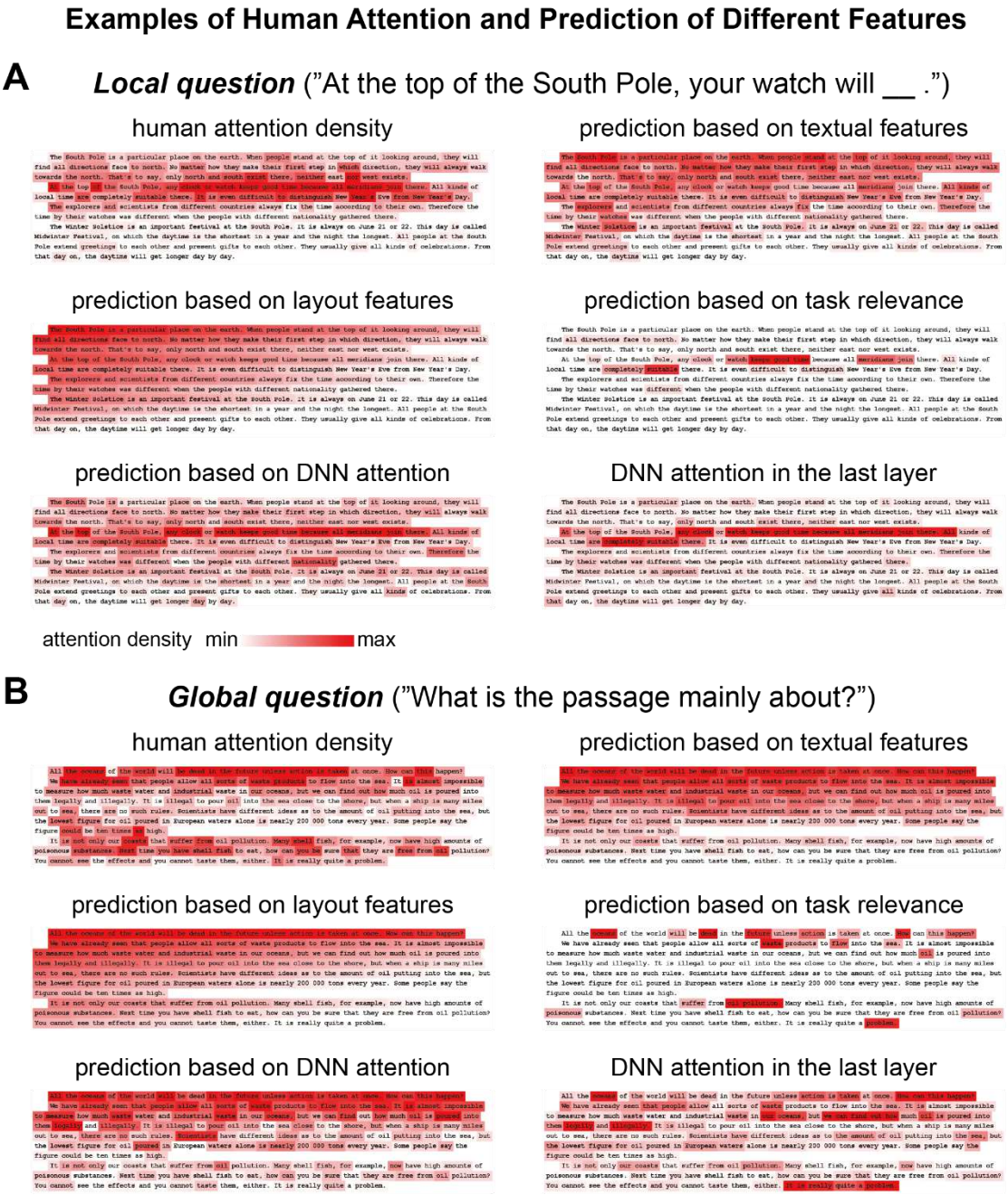
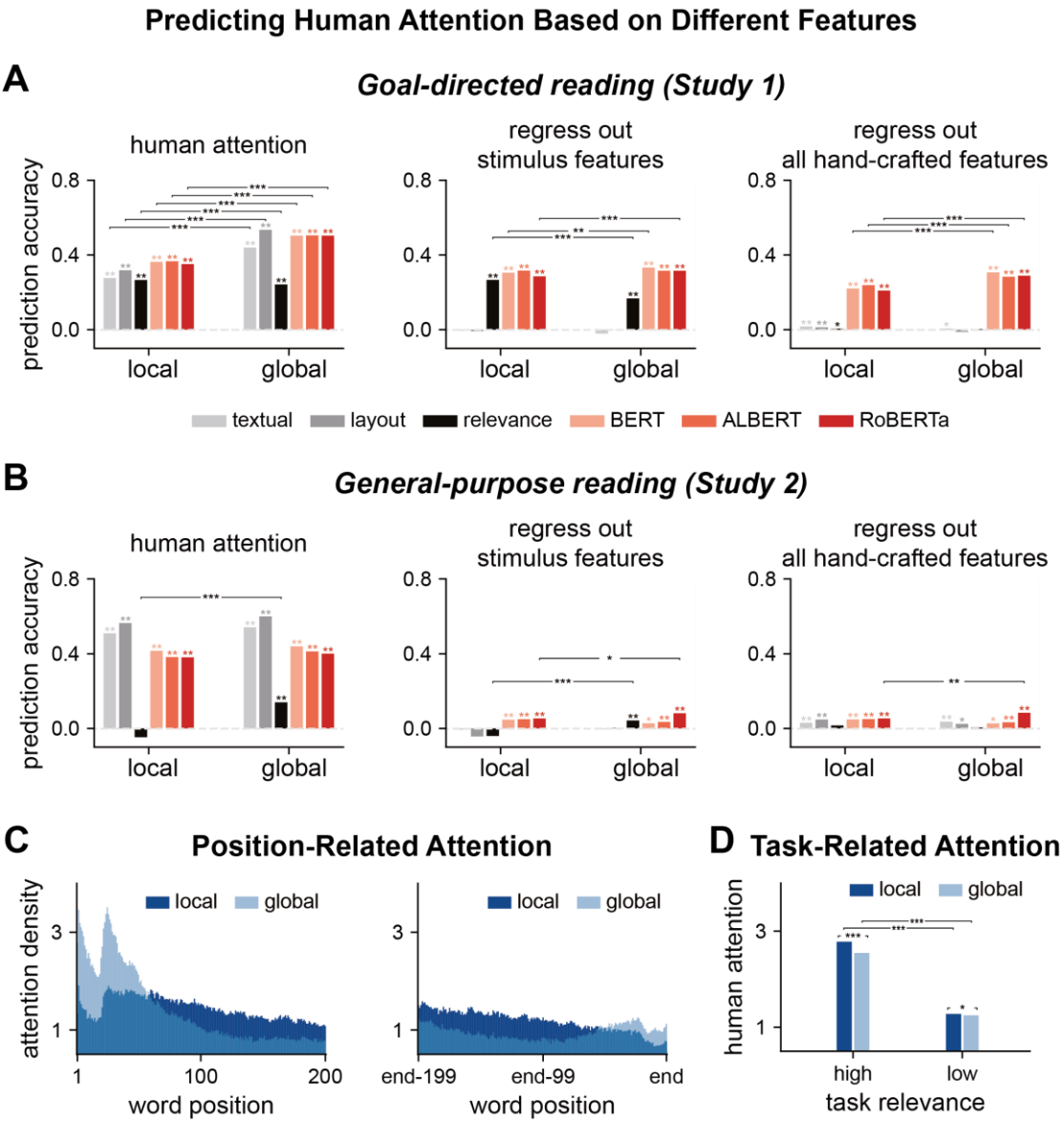


Fig. 2. Examples of human attention distribution and the attention distribution predicted by different features. Panels A and B separately show the attention distribution for two passages and the corresponding questions are shown in the parenthesis. Human attention is quantified by the total fixation time per unit area. Textual features include word properties, e.g., word frequency and the position of a word in the passage. Layout features include visual features that can be processed without recognizing individual words, e.g., line changes. Task relevance refers to human annotation about the contribution of each word to question answering. DNN attention includes all the layers and attention heads, and the DNN attention in the last layer is shown separately (average over attention heads).



195

196

197 **Fig. 3. Predicting human attention using different features.** (**A** and **B**) Panels A
198 and B show the results of Study 1 and Study 2, respectively. The left plots show how
199 well different sets of features can predict human attention. In the middle and right
200 plots, some features are regressed out from human attention, and the residual human
201 attention is predicted by other features. Prediction accuracy that is significantly higher
202 than chance is denoted by stars of the same color as the bar. (**C**) The influence of
203 word position on human attention. Humans generally attend more to the beginning of
204 a passage, especially for global questions. (**D**) The influence of task relevance on
205 human attention. Humans allocate more attention to words that are more relevant to
206 question answering. *P < 0.05; **P < 0.01; ***P < 0.001.

207

Attention Distributions in Humans and DNN

We then investigated whether DNN can generate attention that was comparable to human attention, when optimized to perform the reading task. The general architecture of the models was illustrated in Fig. 1B. The input to the models included all the words in the passage, integrated option, and 3 special tokens. The integrated option was the question concatenated with an option. One of the special token, i.e., CLS, was the decision variable, based on the final representation of which the DNN models decided whether an option was the correct answer or not. In the following, we analyzed the attention weight between the CLS token and each word in the passage (see *Materials and Methods* for details). In each layer of the DNN models, the vectorial representation of the CLS token was updated by a weighted sum of the vectorial representations of all input words and tokens. Therefore, the attention weight on a word could reflect how heavily the word contributed to the decision variable, i.e., the CLS token.

We analyzed 3 DNN models, i.e., BERT[31], ALBERT[32], and RoBERTa[33], and the question answering performance of the 3 DNN models was within the range of human performance (Fig. 1E). Each of the 3 DNN models had 12 layers and each layer had 12 parallel modules that were referred to as attention heads (Fig. 1BC). Each attention head had a separate set of attention weights, and the 12 attention heads were integrated within each layer. Consequently, each word had 144 attention weights (12

layers × 12 heads). In the following, we first tested whether the DNNs could generate human-like attention distributions by attempting to decode human attention distribution from the 144 DNN attention weights using linear regression. Then, we analyzed whether the attention weights in different layers showed different attention properties.

Although the DNN models were only trained to perform the reading comprehension task and were blind to the human fixation data, it was found that the DNN attention weights could significantly predict human attention distribution ($P = 0.002$, permutation test, FDR corrected), and the prediction accuracy was higher for global questions than for local questions ($P = 1.4 \times 10^{-4}$, bootstrap, FDR corrected) (Fig. 3A, left). The prediction accuracy of DNN attention weights was higher than that of textual features and task relevance. Compared with the predictions based on layout features, the predictions based on DNN attention weights were higher for local questions and lower for global questions. It is worth noting that layout features, which were induced by line changes, were not available in the input to DNN models.

DNN attention weights could model the human attention distribution, but did they capture information beyond the hand-crafted features, i.e., textual features, layout features, and task relevance features? We found that when the influences of textual and layout features were regressed out, the residual human attention distribution could still be explained by the DNN attention weights (Fig. 3A, the middle plot). This result

suggested that the DNN attention weights contained information beyond basic stimulus features. Additionally, when the stimulus features and task relevance were both regressed out, the residual human attention distribution remained significantly predicted by the DNN attention weights (Fig. 3A, the right plot). Therefore, DNN attention weights can model human attention and capture information beyond basic hand-crafted features.

Task Modulation in Humans

To further confirm that human attention received top-down modulation from the task, we conducted Study 2 as a control study. In Study 2, participants first read a passage without prior knowledge about the question to answer. After the first-pass passage reading, the participants read the question and were then allowed to read the passage again before answering the question. We analyzed the attention density during the first-pass reading of the passage, which was referred to as general-purpose reading.

For local questions, textual and layout features, but not task relevance, could predict human attention distribution during general-purpose reading (textual features: $P = 0.003$; layout features: 0.003; task relevance: 1; permutation test, FDR corrected). For global questions, all three features could predict human attention distribution ($P = 0.003$ for all 3 features, permutation test, FDR corrected). DNNs could also predict human attention distribution during general-purpose reading, but most of the effect was explained by textual and layout features (Fig. 3B, the middle plot).

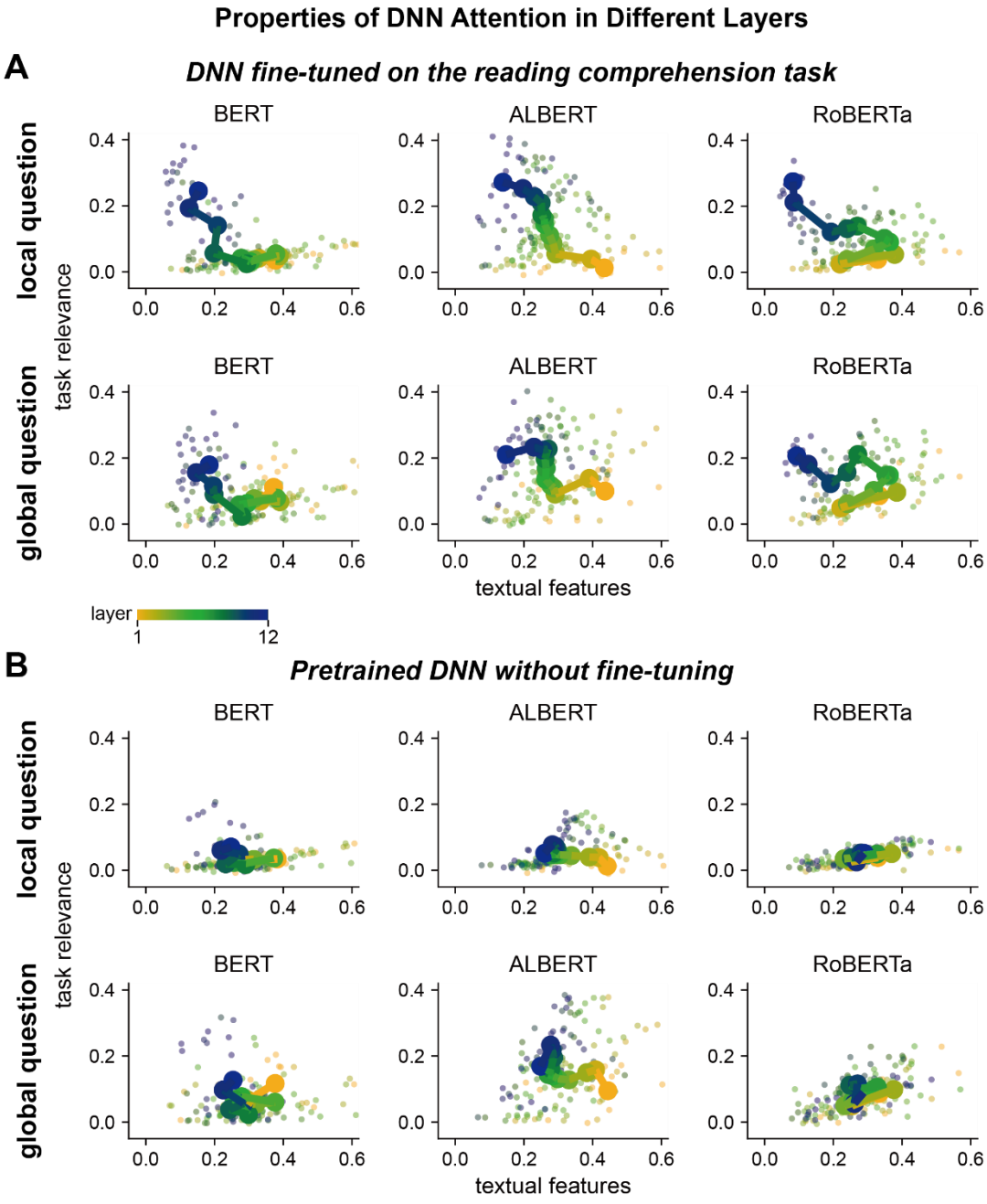


Fig. 4. Influence of stimulus features and top-down task on each DNN layer. The same regression analyses in Fig. 3 are employed to analyze how the DNN attention is affected by lower-level stimulus features and top-down task relevance. Panels A and B show the results for the DNNs fine-tuned based on the reading comprehension task and the pre-trained DNNs that receive no fine-tuning, respectively. Each small dot shows the result from an attention head, and each large dot shows the average over heads of the same layer. Color indicates layer number. Shallow layers of both fine-tuned and pre-trained DNNs are more sensitive to stimulus features. Deep layers of fine-tuned DNN, but not pre-trained DNN, are sensitive to task relevance.

Comparing the results obtained from Study 1 and Study 2, it was evident that human attention could be modulated by the specific reading goal, i.e., the question to answer, on top of textual and layout features. Goal-directed top-down attention, characterized in Study 1, could be modeled by either human-annotated task relevance or the DNN attention weights. In the absence of a specific reading goal, human attention in Study 2 was mainly influenced by stimulus features, e.g., textual and layout features, which were also captured by the DNN attention weights.

DNN Attention in Different Layers

In the human attention system, bottom-up and top-down attention are implemented in separate neural pathways[7,8], and in the following, we analyzed whether attention to stimulus features and attention to task relevant information are also implemented in separate parts of DNN. Since previous studies have shown that different layers in DNNs encoded different types of information[34], we analyzed whether the properties of DNN attention weights differed across layers. Since the layout features were not available to the DNNs, we only considered textual features as stimulus features in this analysis. As shown in Fig. 4A, the attention weights in different layers were sensitive to different features. In general, shallow layers were more strongly influenced by textual features while deeper layers were more strongly influenced by the task relevance. This trend was observed in all 3 DNN models and was especially obvious for local questions. The transitional trajectory across layers, however, was model-dependent in the 2-dimensional feature space. In Fig. 2, examples were shown for the

attention weights averaged over all 12 heads in the last layer of BERT, which resembled the human-annotated task relevance.

Evolution of DNN Attention during Fine-Tuning

All the 3 DNN models were pre-trained based on large-scale corpora and fine-tuned based on the reading comprehension task (Fig. 1D). How was the DNN attention distribution separately influenced by the pre-training and fine-tuning processes? We addressed this question by analyzing the attention weights in pre-trained DNN models that did not receive fine-tuning (Fig. 4B). It was found that the attention weights of pre-trained DNNs were sensitive to textual features in shallow layers but not sensitive to task relevance in deeper layers. This result suggests that top-down attention in DNNs emerged during fine-tuning using the reading comprehension task, while text-based attention existed after pre-training.

We then asked how the attention weights of DNNs changed during fine-tuning and whether such changes were related to the performance of question answering. In the following, we analyzed the properties of models that received different steps of fine-tuning. Furthermore, since fine-tuning process was stochastic, we fine-tuned 10 times (see *Materials and Methods*). We found that, in deep layers, the properties of attention weights significantly changed during fine-tuning (Fig. S3 and Fig. S4). In the last layer, for example, it was clear that the DNN attention weights became more sensitive to task relevance during fine-tuning, coinciding with the improvement in task

329 performance (Fig. 5AD), especially for local questions (Fig. 5A). The trend is less
330 clear for global questions and a potential explanation is that global questions concern
331 the main topic of the passage and can be answered by paying attention to different
332 sets of words. Deep layer's sensitivity to textual features, however, dropped during
333 fine-tuning (Fig. 5BE). Therefore, fine-tuning directed deep layers' attention towards
334 task-relevant information, compromising the sensitivity to textual features.
335 Additionally, we found that the similarity between DNN attention weights and human
336 attention was also boosted by fine-tuning for local but not global questions (Fig.
337 5CF). These results show that fine-tuning boosted task performance by generating
338 goal-directed human-like attention.
339

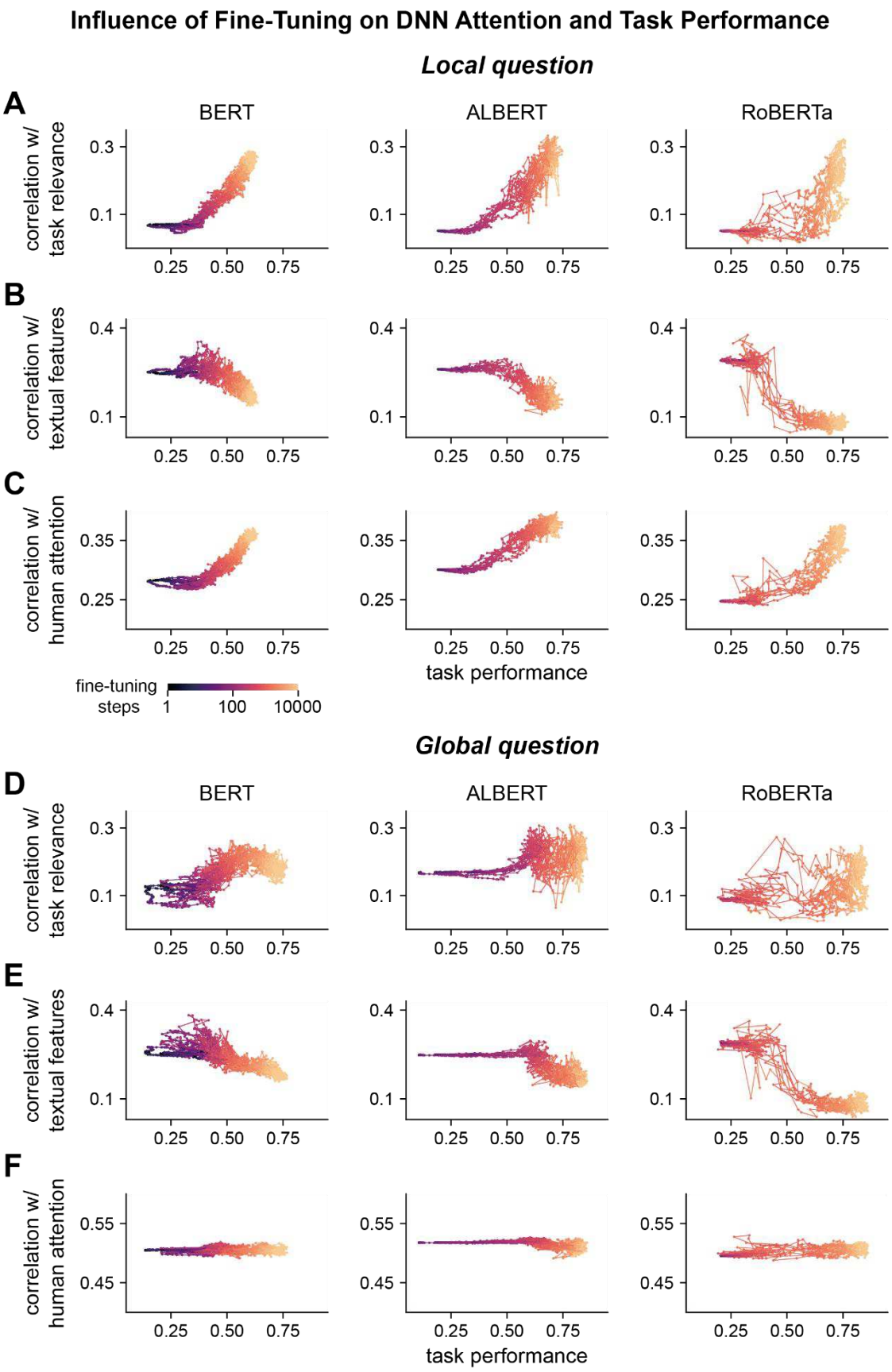

Influence of Fine-Tuning on DNN Attention and Task Performance
Local question
Global question
A
B
C
D
E
F
BERT
ALBERT
RoBERTa
correlation w/ task relevance
correlation w/ textual features
correlation w/ human attention
task performance
fine-tuning steps
1
100
10000

Fig. 5. Influence of fine-tuning on DNN attention and task performance. Each model is fine-tuned 10 times. Each data point denotes the result at a fine-tuning step (color coded), and steps from each run of fine-tuning are connected by a line. (**A**, **B**, **D**, and **E**) The effect of fine-tuning on the attention weights in the last layer of DNNs for local questions (A and B) and global questions (D and E). Fine-tuning enhances the sensitivity to top-down task relevance while reducing the sensitivity to lower-level textual features, which correlates with the increase in task performance. (**C** and **F**) Influence of fine-tuning on the similarity between DNN and human attention. For local questions (C), fine-tuning clearly increases the similarity between DNN and human attention, coinciding with the increase in task performance. For global questions (F), the similarity between DNN and human attention is high even without fine-tuning and is not further boosted by fine-tuning. For ALBERT, 2 out of the 10 runs of fine-tuning are unstable, showing sharp drops in task performance during fine-tuning. Results of these 2 runs are not shown here but separately shown in Fig. S2.

Discussion

The current study quantifies human attention during a real-world reading comprehension task, and demonstrates that DNN models optimized to perform the same task generate human-like attention distribution. Similar to the fact that the human attention system is divided into dual streams separately controlling bottom-up and top-down attention, the shallow and deep layers of DNNs separately implement text-based and goal-directed attention. These results strongly suggest that task optimization provides a computational-level explanation for human attention distribution even in real-world complex reading tasks.

Although rarely investigated, goal-directed reading comprehension provides a suitable paradigm to study attention, and has close relationship to visual search and other information seeking tasks. In the current study, for example, each passage contains a few hundred of words but usually only a few words directly contribute to question answering, highlighting the need for information selection. Furthermore, attributable to the crowding effect[35], the participant could only recognize a couple of words in one fixation and therefore question-relevant key words cannot pop out. Consequently, the participants have to sequentially sample a passage until they could construct their answer. The sampling process, however, is highly active, as evidenced by the uneven distribution of attention. Therefore, goal-directed reading provides a real-world scenario to study attention during complex information seeking behavior[36].

Computational models of attention

Since attention is a key mechanism in the brain, a large number of computational models of attention have been proposed. According to Marr's 3 levels of analysis[37], some models investigate the computational goal of attention. It has been hypothesized that attention can be interpreted as a mechanism to implement optimal learning and decision making. For example, the brain generally attends to information sources that can reduce the uncertainty of a decision[14-16], and a higher degree of attention can enhance the fidelity of neural coding and influence decision[20,38-40]. The current study supports the view that attention is a mechanism to optimize task performance and shows that task optimization can lead to both stimulus-driven bottom-up and goal-directed top-down attention, and can predict human attention in a challenging real-world reading task.

The DNN models analyzed here are trained in two steps (Fig. 1D). The results show that the pre-training process leads to text-based attention while the fine-tuning process leads to goal-directed attention (Fig. 4 and Fig. 5). The pre-training process is also implemented through task optimization. One pre-training task that is used in all 3 models analyzed here is to predict a word based on the context. Another task used to pre-train BERT and ALBERT is to predict the order of two sentences selected from a passage. The purpose of pre-training tasks is to let models learn the general statistical regularity in a language based on large corpora[41], and this process has greatly promoted model performance[31-33]. The results here show that the predictive pre-

405 training tasks can lead to text-based attention. Previous studies have also suggested
406 that predictivity can explain reading time[42,43] and attention to images[44].
407
408 A separate class of models investigate what are the factors contributing to human
409 attention distribution. A large number of models are proposed to predict bottom-up
410 visual saliency[29,30,45], and recently DNN models are also employed to model top-down
411 visual attention. It is shown that, through either implicit[46,47] or explicit training[48,49],
412 DNNs can predict which parts of a picture relates to a verbal phrase, a task similar to
413 goal-directed visual search[50]. The current study distinguishes from these studies in
414 that the DNN model is not trained to predict human attention. Instead, the DNN
415 models naturally generate human-like distribution when trained to perform the same
416 task that humans perform, suggesting that task optimization is a potential cause for the
417 attention distribution during reading.
418
419 **Attention during human reading**
420 How human readers allocate attention during reading is an extensively studied topic.
421 Eye tracking studies have shown that the readers fixate longer at longer words, words
422 of lower-frequency, words that are less predictable based on the context, and words at
423 the beginning of a line, etc[51]. A number of models, e.g., the E-Z reader[52,53] and
424 SWIFT[54], have been proposed to predict the eye movements during reading, either
425 based on basic oculomotor properties or lexical processing[53]. Some models are ideal
426 observer models that minimize the time or the number of saccades required to read a

427 sentence[55,56], but these models do not model how the reading time is affected by the
428 reading goal.
429
430 Traditional eye movement models can generate fine-grained predictions, e.g., which
431 letter in a word will be fixated first, for the reading of simple sentences, but have only
432 been occasionally tested for complex sentences or multi-line text[57]. In contrast to
433 these studies, the current study focuses on goal-directed reading of complex multi-line
434 text and models attention in the units of words. Future studies can potentially integrate
435 classic eye movement models with DNNs to explain the dynamic eye movement
436 trajectory, possibly with a letter-based spatial resolution.
437
438 In most previous studies, readers are instructed to read a sentence in a normal manner,
439 not aimed to extract a specific kind of information[58]. In the current study, however,
440 readers know in advance what question they have to answer and this kind of reading
441 is also referred to as the reading-to-do task[59]. Previous studies have shown the
442 reader's task may have heterogeneous influences on attention, depending on the task
443 difficulty and skill level of readers[60,61]. Here, the task is demanding and the readers
444 are highly skilled to perform the task. Future work is needed to quantify how the task
445 and reading skills modulate human attention and whether these effects can also be
446 modeled by DNN models. Furthermore, when comparing goal-directed and general-
447 purpose reading, it can be observed that attention is more strongly influenced by
448 textual and visual features during general-purpose reading (Fig. 3AB). This result is

consistent with previous results that when the task is to search for a target word, the reading time is much less influenced by, e.g., word frequency compared with normal reading[62].

Attention mechanisms in DNN

In transformer-based DNN models, the roles self-attention plays are highly diverse. Since self-attention assigns a weight between every pair of inputs (including words and special tokens such as CLS), it can capture a number of relationships between words, e.g., co-reference and syntactic dependency[63,64]. In the current study, however, we only analyze a small portion of the self-attention weights that are directly relevant to the task, i.e., the attention weights between CLS and words in the passage. The attention weights analyzed here can be interpreted as a selective information integration mechanism, describing how different words in a passage contribute to the decision variable, i.e., CLS. A couple of recent studies have also compared human attention with several kinds of attention mechanisms in DNNs, and all found some degree of similarity between DNN attention and human attention[65-67]. The current study, however, demonstrate how the task and textual features separately modulate human and DNN attention, and also analyze how the DNN attention properties vary across layers and gradually evolve during training.

Additionally, whether attention can increase the interpretability of DNN models is a heavily debated topic. A number of studies have shown that the DNN attention

471 weights are higher for words that are more important for the task[68,69]. Most of these
472 studies, however, are based on visual inspection of a couple of examples. Other
473 studies, however, reveal low correlation between attention weights and other
474 measures of the importance of words, and therefore raise concern about the
475 interpretablility of attention weights[70,71]. The importance of a word, however, can be
476 measured in many different ways, and no correlation with some importance measures
477 does not indicate no contribution to task solving in other ways. The current study
478 demonstrated that 3 transformer-based models generate human-like attention through
479 task optimization. It remains unclear whether other DNN models show similar
480 properties. Nevertheless, the dataset and methods developed here can be easily
481 applied to test whether other models also evolve human-like attention distribution,
482 serving as a probe to test the biological plausibility of NLP models.
483
484 In sum, the current study demonstrates that, when DNNs and humans perform the
485 same reading comprehension task with comparable accuracy, the DNN attention
486 weights optimized for task performance resemble human attention indexed by eye
487 movements. The results suggest that human attention distribution is shaped by the
488 demand to optimally perform the task. Furthermore, since explanability has become a
489 main issue for DNN models, the results also suggest that, in some models, DNN
490 attention can be interpreted as an approximation of human attention. The large set of
491 eye tracking data in the current study can also motivate future computational

492 modeling of human attention during natural reading tasks and to test whether NLP

493 models exhibit human-like attention distribution.

Materials and Methods

Participants

Study 1 enrolled 102 participants (19-30 years old, mean age, 22.9 years; 54 female). Study 2 enrolled a separate group of 18 participants (21-26 years old, mean age, 23.4 years; 10 female). Additional 11 participants were recruited but failed to pass the calibration process for eye tracking and therefore did not participante in the reading experiments. All participants were native Chinese speakers with normal or corrected-to-normal vision and were college students or graduate students from Zhejiang University, and were thus above the level required to answer high-school-level reading comprehension questions. English proficiency levels were further guaranteed by the following criterion for screening participants: a minimum score of 6 on IELTS, 80 on TOEFL, or 425 on CET6[1]. The experimental procedures were approved by the Research Ethics Committee of the College of Medicine, Zhejiang University (2019–047). The participants provided written consent and were paid.

Experimental materials

The reading materials were selected and adapted from the large-scale RACE dataset, a collection of reading comprehension questions in English exams for middle and high schools in China[23]. We selected eight hundred high-school level questions from the test set of RACE and each question was associated with a distinct passage (117 to 456

[1] The National College English Test (CET) is a national English test system developed to examine the English proficiency of undergraduate students in China. CET includes tests of two levels: a lower level test CET4 and a higher level test CET6.

words per passage). All questions were multiple-choice questions with 4 alternatives including only one correct option among them. The questions fell into 6 types, i.e., Cause (N = 200), Fact (N = 200), Inference (N = 120), Theme (N = 100), Title (N = 100), and Purpose (N = 80). The Cause, Fact, and Inference questions were concerned with the location, extraction, and comprehension of specific information from a passage, and were referred to as local questions. Questions of Theme, Title, and Purpose tested the understanding of a passage as a whole, and were referred to as global questions.

In a separate online experiment, we acquired annotations about the relevance of each word to the question answering task. For each passage, a participant were allowed to annotate up to 5 key words that were considered relevant to answering the corresponding question. Each passage was annotated by N participants (N = 26), producing N versions of annotated key words. Each version of annotation was then validated by a separate participant. In the validation procedure, the participant was required to answer the question solely based on the key words in a specific annotation version; if the person could not derive the correct answer, this version of annotation was discarded. The percent of questions correctly answered in the validation procedure was 75.9% and 67.6%, for local and global questions respectively. The correct rate for local questions was comparable to the correct rate in the eye tracking experiments, in which the readers had access to the whole passage (Fig. 1), suggesting that as few as 5 key words could well support question answering. The correct rate for

global questions was lower than the eye tracking experiments but remained much higher than chance, i.e., 25%. For the *M* versions of annotation that passed the validation procedure, if a word was annotated in *K* versions, the relevance of the word to question answering, i.e., task relevance, was *M/K*. More details about the question types and the annotation procedures could be found in reference[72].

Experimental procedures

Study 1: Study 1 included all 800 passages, and different question types were separately tested in different experiments, hence six experiments in total. Each experiment included 25 participants and one participant could participate in multiple experiments. Before each experiment, to familiarize the participants with the experimental procedure and the type of questions to answer, participants were given a familiarization session with 5 questions that were not used in the formal experiment or in the analysis. During the formal experiment, questions were presented in a randomized order. Considering the quantities of questions, for Cause and Fact questions, the experiment was carried out in 3 separate days (one third questions on each day), and for other question types the experiment was carried out in 2 days (fifty percent of questions on each day).

The experiment procedure in Study 1 was illustrated in Fig. 1A. In each trial, participants first read a question, pressed the space bar to read the corresponding passage, and then pressed it again to read the question coupled with 4 options and

answer the question. The time limit for passage reading was 120 s. To encourage the participants to read as quickly as possible, the bonus they received for a specific question would decrease linearly from 1.5 to 0.5 RMB over time. They did not receive any bonus for the question, however, if they gave a wrong answer. Furthermore, before answering the comprehension question, the participants reported whether they were confident that they could correctly answer the question (yes or no). Participants selected yes for 90.47% of questions (89.62% and 92.04% for local and global questions, respectively). After answering the question, they also rated their confidence about their answer on the scale of 1-4 (low to high). The mean confidence rating was 3.25 (3.28 and 3.18 for local and global question, respectively), suggesting that the participants were confident about their answers.

Study 2: Study 2 included 96 reading passages and questions, with 16 questions for each question type that were randomly selected from the questions used in Study 1. The study was carried out in 2 days, and none of the participants participated in Study 1. The familiarization procedure was identical to that in Study 1.

The procedure of Study 2 was similar to that of Study 1, except a difference that a 90-s first-pass passage reading stage was introduced at the beginning of each trial. During the first-pass passage reading, participants had no prior information of the relevant question. The participants could press the space bar to terminate the first-pass reading stage and to read a question. Then, participants read the passage for the second time

with a time limit of 30 s, before proceeding to answer the question. The payment method was similar to Study 2, and the bonus was calculated based on the duration of second-pass passage reading.

Stimulus presentation and eye tracking

The text was presented using the bold Courier New font, and each letter occupied 14 × 27 pixels. We set the maximum number of letters on each line to 120 and used double space. We separated paragraphs by indenting the first line of each new paragraph. Participants sat about 880 mm from a monitor, at which each letter horizontally subtended approximately 0.25 degrees of visual angle.

Eye tracking data were recorded from the left eye with 500-Hz sampling rate (Eyelink Portable Duo, SR Research). The experiment stimuli were presented on a 24-inch monitor (1920 × 1080 resolution; 60 Hz refresh rate) and administered using MATLAB Psychtoobox[73]. Each experiment started with a 13-point calibration and validation of eye tracker, and the validation error was required to be below 0.5º of visual angle. Furthermore, before each trial, a 1-point validation was applied, and if the calibration error was higher than 0.5º, a recalibration was carried out. Head movements were minimized using a chin and forehead rest.

DNN models

We tested 3 popular transformer-based DNN models, i.e., BERT[31], ALBERT[32], and

RoBERTa[33]. ALBERT and RoBERTa were both adapted from BERT, and had the same basic structure. RoBERTa differed from BERT in its pre-training procedure[33] while ALBERT applied factorized embedding parameterization and cross-layer parameter sharing to reduce memory consumption[32]. Following previous works[32,33], each option was independently processed. For the i^{th} option (i = 1, 2, 3, or 4), the question and the option were concatenated to form an integrated option. As shown in Fig. 1B, for the i^{th} option, the input to DNNs was the following sequence:

$$C_i, P_1, P_2, ..., P_N, S_{i,1}, O_{i,1}, O_{i,2}, ..., O_{i,M}, S_{i,2},$$

where C_i, $S_{i,1}$, and $S_{i,2}$ denoted special tokens, i.e., the CLS, SEP_1, and SEP_2 tokens, separating different components of the input. P_1, P_2, …, P_N denoted all the N words of a passage, while $O_{i,1}$, $O_{i,2}$, …, $O_{i,M}$ denoted all the M words of the i^{th} integrated option. Each of the token was represented by a vector. The vectorial representation was updated in each layer, and in the following the output of the l^{th} layer was denoted as a superscript, e.g., C_i^l. Following previous works[32,33], we calculated a score for each option, which indicated the possibility that the option was the correct answer. The score was calculated by first applying a linear transform to the final representation of the CLS token, i.e.,

$$s_i = \Phi C_i^{12},$$

where C_i^{12} was the final output representation of CLS and Φ was a vector learned

624 from data. The score was independently calculated for each option and then
625 normalized using the following equation:
626
627 $$score_i = \frac{exp(s_i)}{\sum_{i=1}^{4} exp(s_i)}.$$
628 The answer to a question was determined as the option with highest score, and all the
629 models were trained to maximize the logarithmic score of the correct option.
630
631 We fine-tuned the DNNs based on the training set of RACE. In the analyses shown in
632 Fig. 5, the fine-tuning process was independently run 10 times. Each time, the training
633 samples were fed in with a randomized order and nodes in the dropout layer were
634 randomly eliminated. Results from the first run of fine-tuning was used for the main
635 analysis reported in Figs. 2-4. All models were implemented based on HuggingFace[74]
636 and all hyperparameters for fine-tuning were adopted from previous studies (Table
637 S1).
638
639 To isolate how the fine-tuning process modulated DNN attention, we also tested the
640 pre-trained DNN that was not fine-tuned on RACE dataset, and compared it with the
641 fine-tuned model (Fig. 4). Furthermore, we quantified how the properties of DNN
642 attention changed throughout the fine-tuning process by analyzing models that
643 received different steps of fine tuning. The steps we sampled were exponentially
644 spaced between 1 and the maximum fine-tuning steps.
645

DNN attention

In each attention head, the attention mechanism calculated an attention weight between any pair of inputs, including words and special tokens. The vectorial representation of each input was then updated by the weighted sum of the vectorial representations of all inputs[22]. In other words, the models we considered were all context-dependent models, in which the representation of each word was modeled by integrating the representations of all inputs. Since only the CLS token was directly related to question answering, here we analyzed the attention weights that were used to calculate the vectorial representation of CLS (illustrated in Fig. 1C). For each layer, the output of an attention head, i.e., C^h, was computed using the following equations. For the sake of clarity, we denote the input words and tokens generally as X_i.

$$C^h = \sum_{i=1}^{N+M+2} \alpha_i V_i = \alpha_C V_C + \sum_{n=1}^{N} \alpha_{Pn} V_{Pn} + \alpha_{S1} V_{S1} + \sum_{m=1}^{M} \alpha_{Om} V_{Om} + \alpha_{S2} V_{S2},$$

$$\alpha_i = \frac{exp(Q_C K_i^{\mathrm{T}})}{\sum_{i=1}^{N+M+2} exp(Q_C K_i^{\mathrm{T}})},$$

$$V_i = X_i W^V + b^V,\ K_i = X_i W^K + b^K,\ Q_C = X_C W^Q + b^Q,$$

where W^v, W^q, W^k, b^v, b^q, and b^k were parameters to learn from the data, and α_i was the attention weight between CLS and X_i. The attention weight between CLS and the n^{th} word in the passage, i.e., α_{Pn}, was compared to human attention. Here, we only considered the attention weights associated with the correct option.

667 Output of the attention module, i.e., C^h, was concatenated over all the 12 heads in
668 each layer, and further processed by position-wise operations to generate the final
669 representation of CLS in the layer[22]. Additionally, DNNs used byte-pair tokenization
670 which split some words into multiple tokens. We converted the token-level attention
671 weights to word-level attention weights by summing the attention weights over tokens
672 within a word[63,66].
673
674 **Human attention analysis and prediction**
675 We analyzed eye fixations during passage reading in Study 1 and the first-pass
676 passage reading in Study 2. The total fixation time of each word was extracted using
677 the SR Research Data Viewer softwave. We averaged the total fixation time across all
678 participants who correctly answered the question, and measured human attention
679 using the attention density, i.e., the total fixation time divided by the area a word
680 occupied.
681
682 We employed linear regression to test whether a set of features could explain human
683 attention distribution. Four sets of features were analyzed, i.e., textual features, layout
684 features, task relevance, and DNN attention weights. The textual features included
685 word length, logarithmic word frequency estimated based on the British National
686 Corpus[75], ordinal position of a word in a sentence, ordinal position of a word in a
687 passage, and ordinal sentence number of a word. The layout features referred to the
688 visual layout of text, i.e., features induced by line changes, including the coordinate of

689 the left most pixel of a word, ordinal position of a word in a paragraph, ordinal row
690 number of a word in a paragraph, ordinal row number of a word in a passage. Task
691 relevance was annotated by humans (see *Experimental materials* for details), and the
692 DNN attention weights included the 144 attention weights from all layers and
693 attention heads. In the regression analysis, human attention density on word *w* was
694 modeled using the following equation.
695
696 $$attention\_density_w = \sum_{j=1}^{J} \beta_j F_{w,j} + b + \varepsilon_w,$$
697
698 where F and ε denoted the features being considered and the residual error,
699 respectively. The parameters β and b were fitted to minimize the mean square error.
700 Each feature and the human attention distribution were normalized within a passage
701 by taking the z-score. The prediction accuracy, i.e., the correlation between predicted
702 attention and actual human attention, was calculated based on five-fold cross-
703 validation. Each question type was separately modeled.
704
705 **Statistical tests**
706 We employed a one-sided permutation test to test whether a set of features were
707 statistically significant to predict human attention. Five hundred chance-level
708 prediction accuracy was calculated by predicting shuffled human attention.
709 Specifically, the human attention density was shuffled across words and was predicted
710 by word features which were not shuffled. The procedure was repeated 500 times,

creating 500 chance-level prediction accuracy. If the actual correlation was greater than N out of the 500 chance-level correlation, the significance level was $(N + 1)/501$.

The comparison between global and local questions were based on bias-corrected and accelerated bootstrap[76]. For example, to test whether the prediction accuracy differed between the 2 types of questions, all global questions were resampled with replacement 50000 times and each time the prediction accuracy was calculated based on the resampled questions, resulting in 50000 resampled prediction accuracy. If the prediction accuracy for local questions was greater (or smaller) than N out of the 50000 resampled accuracy for global questions, the significance level of their difference was $2(N + 1)/50001$. When multiple comparisons were performed, the p-value was further adjusted using the false discovery rate (FDR) correction.

Acknowledgements

We thanks David Poeppel and Erik D. Reichle for valuable comments on earlier versions of this manuscript; Jonathan Simon, Bingjiang Lyu, and members of the Ding lab for thoughtful discussions and feedback; Yuran Zhang, Anqi Dai, Zhonghua Tang, and Yuhan Lu for assistance with experiments. Work supported by National Natural Science Foundation of China 31771248 and Major Scientific Research Project of Zhejiang Lab 2019KB0AC02

Author contributions

Nai Ding acquired the funding, conceived and coordinated the project, analyzed data, and wrote the manuscript. Jiajie Zou implemented the experiments and models, analyzed data, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

References

1 Posner, M. I. & Petersen, S. E. The attention system of the human brain. *Annu. Rev. Neurosci.* **13**, 25-42 (1990).

2 Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97-136 (1980).

3 Posner, M. I. & Boies, S. J. Components of attention. *Psychol. Rev.* **78**, 391 (1971).

4 Chun, M. M., Golomb, J. D. & Turk-Browne, N. B. A taxonomy of external and internal attention. *Annu. Rev. Psychol.* **62**, 73-101 (2011).

5 Awh, E., Belopolsky, A. V. & Theeuwes, J. Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends Cogn. Sci.* **16**, 437-443 (2012).

6 Egeth, H. E. & Yantis, S. Visual attention: Control, representation, and time course. *Annu. Rev. Psychol.* **48**, 269-297 (1997).

7 Corbetta, M. & Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **3**, 201-215 (2002).

8 Buschman, T. J. & Miller, E. K. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* **315**, 1860-1862 (2007).

9 Knudsen, E. I. Fundamental components of attention. *Annu. Rev. Neurosci.* **30**, 57-78 (2007).

10 Atiani, S. *et al.* Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. *Neuron* **82**, 486-499 (2014).

11 Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193-222 (1995).

12 Broadbent, D. Perception and communication Pergamon. *Elmsford, New York* (1958).

13 Kahneman, D. *Attention and effort*. Vol. 1063 (Citeseer, 1973).

14 Dayan, P., Kakade, S. & Montague, P. R. Learning and selective attention. *Nat. Neurosci.* **3**, 1218-1223 (2000).

15 Gottlieb, J., Hayhoe, M., Hikosaka, O. & Rangel, A. Attention, reward, and information seeking. *J. Neurosci.* **34**, 15497-15504 (2014).

16 Yu, A. J. & Dayan, P. Inference, attention, and decision in a Bayesian neural architecture. In *Proc. Advances in neural information processing systems* 1577-1584 (MIT Press, 2005).

17 Radulescu, A., Niv, Y. & Ballard, I. Holistic reinforcement learning: the role of structure and attention. *Trends Cogn. Sci.* **23**, 278-292 (2019).

18 Najemnik, J. & Geisler, W. S. Optimal eye movement strategies in visual search. *Nature* **434**, 387-391 (2005).

19 Navalpakkam, V., Koch, C., Rangel, A. & Perona, P. Optimal reward harvesting in complex perceptual environments. *Proc. Natl Acad. Sci. USA* **107**, 5232-5237 (2010).

20 Yang, S. C.-H., Lengyel, M. & Wolpert, D. M. Active sensing in the categorization of visual patterns. *Elife* **5**, e12215 (2016).

21 Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proc. 3rd International Conference on Learning Representations* (ICLR, 2015).

22 Vaswani, A. *et al.* Attention is all you need. In *Proc. Advances in neural information processing systems* 5998-6008 (Curran Associates, 2017).

23 Lai, G., Xie, Q., Liu, H., Yang, Y. & Hovy, E. Race: Large-scale reading comprehension dataset from examinations. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing* 785-794 (Association for Computational Linguistics, 2017).

24 Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619-8624 (2014).

25 Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630-644. e616 (2018).

26 Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55-67 (2021).

27 White, S., Chen, J. & Forsyth, B. Reading-related literacy activities of American adults: Time spent, task types, and cognitive skills used. *J. Lit. Res.* **42**, 276-307 (2010).

28 Rayner, K. & McConkie, G. W. What guides a reader's eye movements? *Vision Res.* **16**, 829-837 (1976).

29 Borji, A., Sihite, D. N. & Itti, L. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Trans. Image Process.* **22**, 55-69 (2013).

30 Bylinskii, Z., Judd, T., Oliva, A., Torralba, A. & Durand, F. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 740-757 (2018).

31 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4171-4186 (Association for Computational Linguistics, 2019).

32 Lan, Z. *et al.* Albert: A lite bert for self-supervised learning of language representations. In *Proc. International Conference on Learning Representations* (ICLR, 2020).

33 Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).

34 Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Object detectors emerge in deep scene CNNs. In *Proc. International Conference on Learning Representations* (ICLR, 2015).

35 Pelli, D. G. *et al.* Crowding and eccentricity determine reading rate. *J. Vis.* **7**, 20-20 (2007).

36 Gottlieb, J., Cohanpour, M., Li, Y., Singletary, N. & Zabeh, E. Curiosity, information demand and attentional priority. *Curr. Opin. Behav.* **35**, 83-91 (2020).

37 Marr, D. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY* **2** (1982).

38 Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292-1298 (2010).

39 Jang, A. I., Sharma, R. & Drugowitsch, J. Optimal policy for attention-modulated decisions explains human fixation behavior. *Elife* **10**, e63436 (2021).

40 Wyart, V., Myers, N. E. & Summerfield, C. Neural mechanisms of human perceptual choice under focused and divided attention. *J. Neurosci.* **35**, 3485-3498 (2015).

41 Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. (2018).

42 Hale, J. Information‐theoretical complexity metrics. *Lang. Linguist. Compass.* **10**, 397-412 (2016).

43 Smith, N. J. & Levy, R. The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302-319, doi:10.1016/j.cognition.2013.02.013 (2013).

44 Itti, L. & Baldi, P. Bayesian surprise attracts human attention. *Vision Res.* **49**, 1295-1306 (2009).

45 Tatler, B. W., Hayhoe, M. M., Land, M. F. & Ballard, D. H. Eye guidance in natural vision: reinterpreting salience. *J. Vis.* **11**, 5-25, doi:10.1167/11.5.5 (2011).

46 Anderson, P. *et al.* Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. The IEEE Conference on Computer Vision and Pattern Recognition* 6077-6086 (CVPR, 2018).

47 Xu, K. *et al.* Show, attend and tell: Neural image caption generation with visual attention. In *Proc. 32nd International Conference on Machine Learning* 2048-2057 (PMLR, 2015).

48 Das, A., Agrawal, H., Zitnick, L., Parikh, D. & Batra, D. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Comput. Vis. Image Underst.* **163**, 90-100 (2017).

49 Liu, C., Mao, J., Sha, F. & Yuille, A. Attention correctness in neural image captioning. In *Proc. AAAI Conference on Artificial Intelligence* 4176-4182 (AAAI, 2017).

50 Wolfe, J. M. & Horowitz, T. S. Five factors that guide attention in visual search. *Nat. Hum. Behav.* **1**, 1-8 (2017).

51 Just, M. A. & Carpenter, P. A. A theory of reading: From eye fixations to comprehension. *Psychol. Rev.* **87**, 329-354 (1980).

52 Reichle, E. D., Pollatsek, A., Fisher, D. L. & Rayner, K. Toward a model of eye movement control in reading. *Psychol. Rev.* **105**, 125-157 (1998).

53 Reichle, E. D., Rayner, K. & Pollatsek, A. The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behav. Brain Sci.* **26**, 445-476 (2003).

54 Engbert, R., Nuthmann, A., Richter, E. M. & Kliegl, R. SWIFT: a dynamical model of saccade generation during reading. *Psychol. Rev.* **112**, 777-813 (2005).

55 Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S. & Tjan, B. S. Mr. Chips 2002: New insights from an ideal-observer model of reading. *Vision Res.* **42**, 2219-2234 (2002).

56 Liu, Y. & Reichle, E. The emergence of adaptive eye movements in reading. In *Proc. Annual Meeting of the Cognitive Science Society*.

57 Mancheva, L. *et al.* An analysis of reading skill development using EZ Reader. *J. Cogn. Psychol.* **27**, 657-676 (2015).

58 Clifton, C. *et al.* Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *J. Mem. Lang.* **86**, 1-19, doi:10.1016/j.jml.2015.07.004 (2016).

59 Duffy, T. M. & Kabance, P. Testing a readable writing approach to text revision. *J. Educ. Psychol.* **74**, 733-748 (1982).

60 van der Schoot, M., Vasbinder, A. L., Horsley, T. M. & van Lieshout, E. C. D. M. The role of two reading strategies in text comprehension: An eye fixation study in primary school children. *J. Res. Read.* **31**, 203-223 (2008).

61 Kaakinen, J. K., Hyönä, J. & Keenan, J. M. How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *J. Exp. Psychol.: Learn. Mem. Cogn.* **29**, 447-457 (2003).

62 Rayner, K. & Fischer, M. H. Mindless reading revisited: Eye movements during reading and scanning are different. *Perception & Psychophysics* **58**,

734-747 (1996).

63 Clark, K., Khandelwal, U., Levy, O. & Manning, C. D. What Does BERT Look at? An Analysis of BERT's Attention. In *Proc. 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* 276-286 (Association for Computational Linguistics, 2019).

64 Voita, E., Talbot, D., Moiseev, F., Sennrich, R. & Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 5797-5808 (Association for Computational Linguistics, 2019).

65 Sood, E., Tannert, S., Frassinelli, D., Bulling, A. & Vu, N. T. Interpreting attention models with human visual attention in machine reading comprehension. In *Proc. 24th Conference on Computational Natural Language Learning* 12–25 (Association for Computational Linguistics).

66 Bolotova, V. *et al.* Do People and Neural Nets Pay Attention to the Same Words: Studying Eye-tracking Data for Non-factoid QA Evaluation. In *Proc. 29th ACM International Conference on Information & Knowledge Management* 85-94 (ACM, 2020).

67 Sen, C., Hartvigsen, T., Yin, B., Kong, X. & Rundensteiner, E. Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 4596-4608.

68 Yang, Z. *et al.* Hierarchical attention networks for document classification. In *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1480-1489 (Association for Computational Linguistics, 2016).

69 Lin, Z. *et al.* A structured self-attentive sentence embedding. In *Proc. International Conference on Learning Representations* (ICLR, 2017).

70 Serrano, S. & Smith, N. A. Is Attention Interpretable? In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 2931-2951 (Association for Computational Linguistics, 2019).

71 Jain, S. & Wallace, B. C. Attention is not Explanation. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics* 3543-3556 (Association for Computational

Linguistics).

72 Zou, J. *et al.* PALRACE: Reading Comprehension Dataset with Human Data and Labeled Rationales. Preprint at https://arxiv.org/abs/2106.12373 (2021).

73 Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433-436 (1997).

74 Wolf, T. *et al.* HuggingFace's Transformers: State-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38-45 (Association for Computational Linguistics, 2020).

75 Burnard, L. *The British National Corpus, version 3 (BNC XML Edition).* <http://www.natcorp.ox.ac.uk/> (2007).

76 Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap*. (CRC press, 1994).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SM1030.pdf