Deep Neural Networks Evolve Human-like Attention Distribution during Goal-directed Reading Comprehension

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

Article

Keywords: deep neural networks, reading comprehension

Posted Date: August 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-813994/v1

Deep Neural Networks Evolve Human-like Attention Distribution during Goal-directed Reading Comprehension

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

[1]Zhejiang lab; College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou 311121, China

[2]Key Laboratory for Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou 310027, China

***Corresponding author:**

Nai Ding,

Email: ding_nai@zju.edu.cn

Zhejiang lab; College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou 311121, China

Abstract

Attention is a key mechanism for information selection in both biological brains and many state-of-the-art deep neural networks (DNNs). Here, we investigate whether humans and DNNs allocate attention in comparable ways when seeking information in a text passage to answer a question. We analyze 3 transformer-based DNNs that reach human-level performance when trained to perform the reading comprehension task. We find that the DNN attention distribution quantitatively resembles human attention distribution measured by eye tracking: Human readers fixate longer on words that are more relevant to the question-answering task, demonstrating that attention is modulated by the top-down reading goal, on top of lower-level visual layout and textual features. Further analyses reveal that the attention weights in DNNs are also influenced by both the top-down reading goal and lower-level textual features, with the shallow layers more strongly influenced by lower-level textual features and the deep layers attending more to task-relevant words. Additionally, deep layers' attention to task-relevant words gradually emerges when pre-trained DNN models are fine-tuned to perform the reading comprehension task, which coincides with the improvement in task performance. These results demonstrate that DNNs can naturally evolve human-like attention distribution through task optimization. The results suggest that human attention during goal-directed reading comprehension is a consequence of task optimization and the attention weights in DNN are of biological significance.

Introduction

Artificial intelligence (AI) and cognitive science separately investigate how machines and brains solve complex information processing problems, such as language comprehension and visual object recognition. As the artificial neural network approach in AI was in part inspired by biological neural networks, there is continuing interest in comparing the performance of artificial and biological neural networks[1-11]. If human- or animal-like behaviors emerge in artificial neural networks, it indicates that the computations implemented in artificial neural networks can serve as a possible model for human/animal cognition[12-15]. Indeed, it has been found that artificial neurons in modern DNNs can evolve receptive field properties that are comparable to those measured from animal visual cortices[16], and DNN models that have properties more consistent with biological neural systems tend to perform better at information processing tasks[10,11].

Traditional artificial neural networks mainly mimic lower-level or biophysical properties of neurons, while the new generations of DNN models also attempt to mimic high-level cognitive functions, e.g., attention. Attention mechanisms have greatly improved the performance of DNNs and have become a necessary component in state-of-the-art DNN models, especially in the field of natural language processing (NLP)[17-21]. Recent studies have shown that the attention mechanism in DNN can play a wide variety of roles in language processing, e.g., to extract task relevant information[22,23] and to analyze syntactic dependencies and semantic co-reference[24-26]. The attention mechanism in DNNs, however, is not designed to quantitatively simulate human

attention, and few studies have systematically compared human and DNN attention during the same language processing task (see Bolotova et al. 2020[27] for a notable exception). Therefore, it remains unclear to what extent the attention mechanisms in DNN language models are comparable to human attention and whether the attention mechanisms in DNN can serve as a model for human attention.

The human attention system has multiple components which contribute differently to different tasks. For example, when freely viewing an image, attention is primarily modulated by visual saliency, and this kind of attention is referred to as bottom-up attention[28,29]. When searching for a target object in a visual scene, however, viewers attend more to possible locations of the target and objects with visual features more consistent with the target[30]. This form of attention - induced by the task - is called top-down attention[28,29]. In visual perception tasks that mainly engage bottom-up attention, a large number of studies have shown that neural networks can be trained to model human attention distribution measured through eye tracking[31,32]. Recently, some models have also been proposed to model top-down attention[9,33-36]. In the domain of language processing, computational models have been proposed to predict human readers' eye movements when they read simple sentences without a specific purpose[37], a task similar to free viewing. To our knowledge, however, no model has been proposed to predict human attention when readers read a passage with a specific goal, e.g., to answer a question, although goal-directed reading was the most common reading behavior for adults[38].

Here we compare the attention distribution for humans and DNNs during a reading comprehension task in which humans or DNNs have to answer a question by reading a passage. We select this task since it is a benchmark task to test NLP algorithms[39,40] and also a common task to test human verbal ability, e.g., in exams such as SAT, GRE, and TOEFL. This task is also suitable to investigate attention, since a passage contains an enormous amount of information, but only a small portion of it is typically relevant to answering a specific question, imposing a strong load for information selection. Finally, state-of-the-art DNN models have recently achieved human-level performance on the reading comprehension task for questions at the difficulty level corresponding to high school exams in China[18,19,21].

With the reading comprehension task, we quantified the attention mechanisms in humans using the fixation time and the attention mechanisms in DNNs using the attention weight on each word. Note that both human eye fixations and the attention weights in DNNs reflect intermediate processing steps instead of the outcome of reading comprehension. We aim to investigate three closely related questions by analyzing and comparing human and DNN attention. First, how is human and DNN attention modulated by stimulus features and the top-down reading goal, i.e., the need to answer a specific question? Second, do humans and DNNs show similar attention distribution? Third, how does the DNN attention distribution evolve during training and how does it relate to task performance?

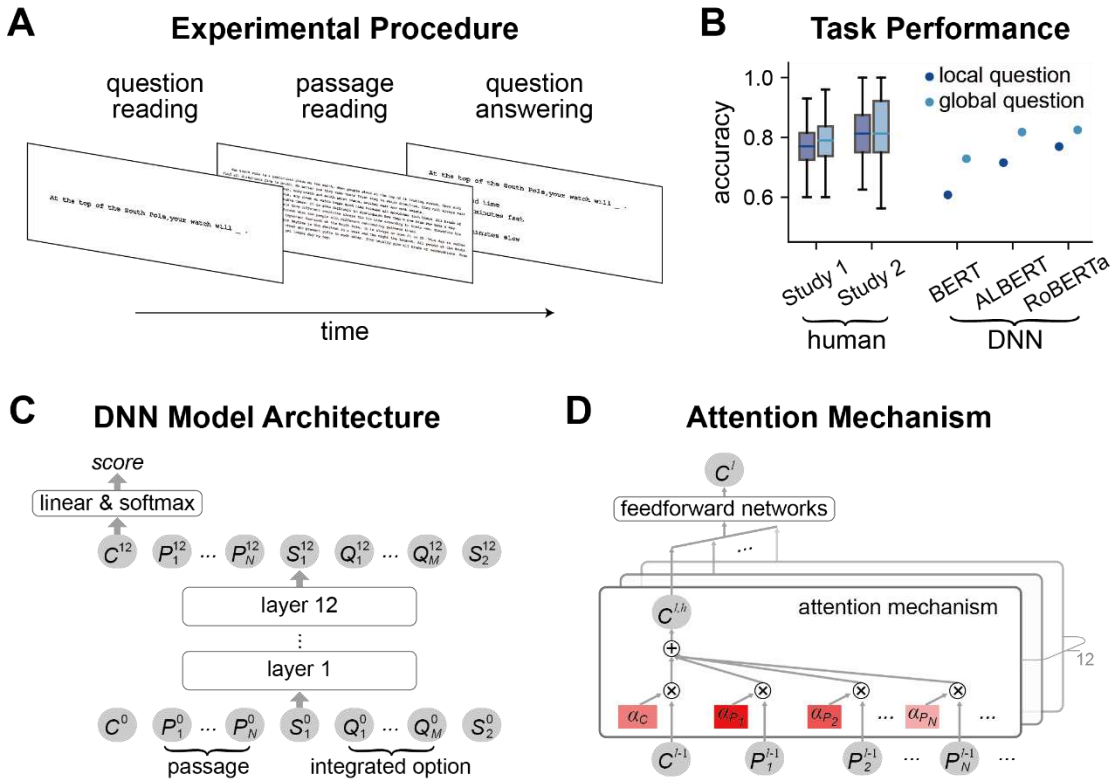

Fig. 1. Experimental procedure and DNN model. (**A**)The experimental procedure in Study 1. In each trial, participants read a question first, and then read the corresponding passage, and finally proceed to read the question, coupled with 4 options, and answer it. (**B**) Performance of humans and DNN models on the reading comprehension task. (**C**) 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). The CLS token integrates information across words and is used to calculate a score that reflects how likely the option is the correct answer. The DNN model has 12 layers and has 12 attention heads in each layer. (**D**) 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. For example, the vectorial representation of the CLS token is a weighted sum of the vectorial representations of all words and tokens. 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.

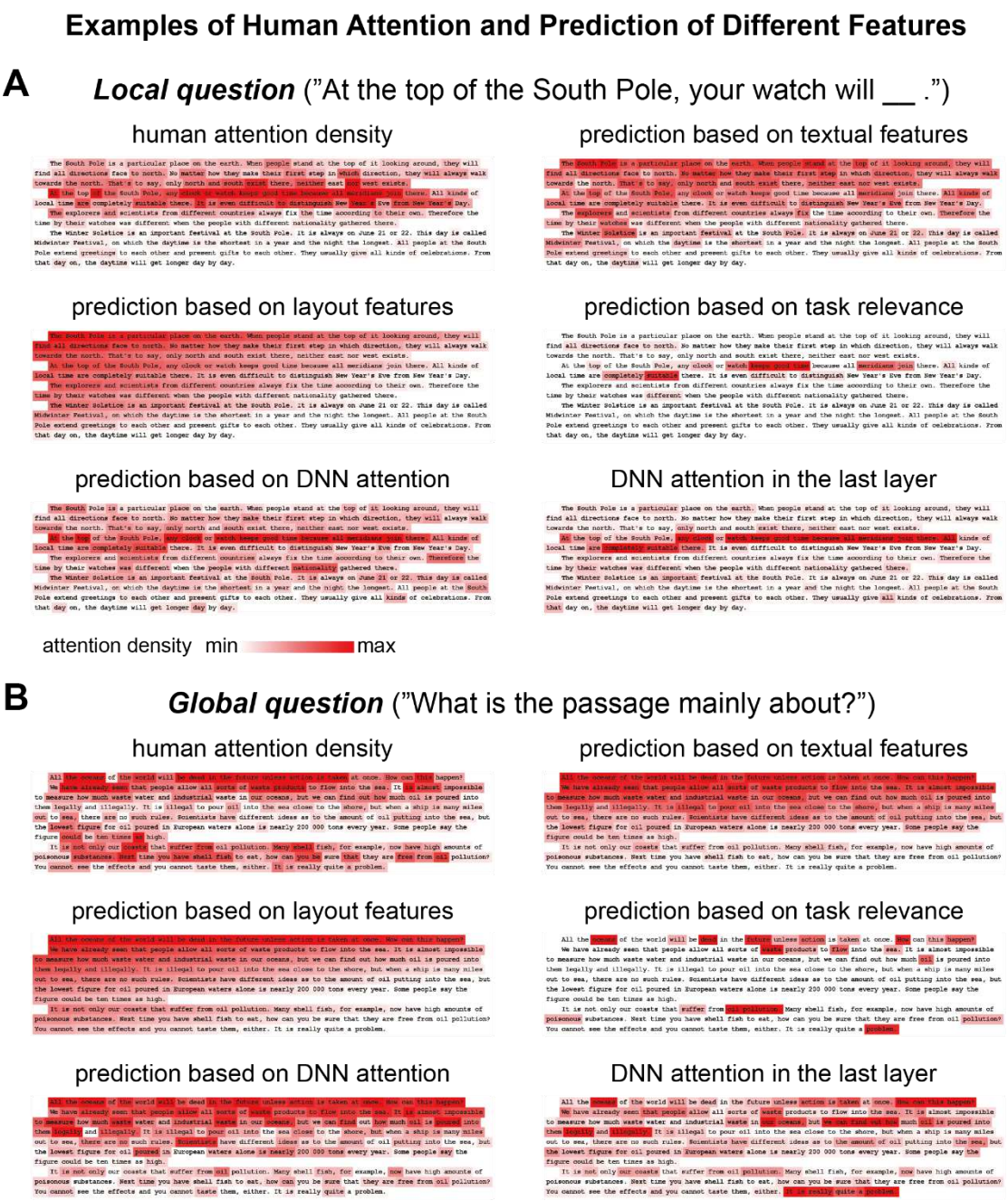

Fig. 2. Examples of the human attention distribution and the human attention distribution predicted by different features. Panels A and B separately show the attention distribution for two passages and the corresponding questions are shown in the parenthesis. Human attention is quantified by the total 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. Task relevance contains 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 (averaged over attention heads).

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. 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. 1B, 77.49% and 78.77% for local and global questions, respectively).

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[41]. 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.

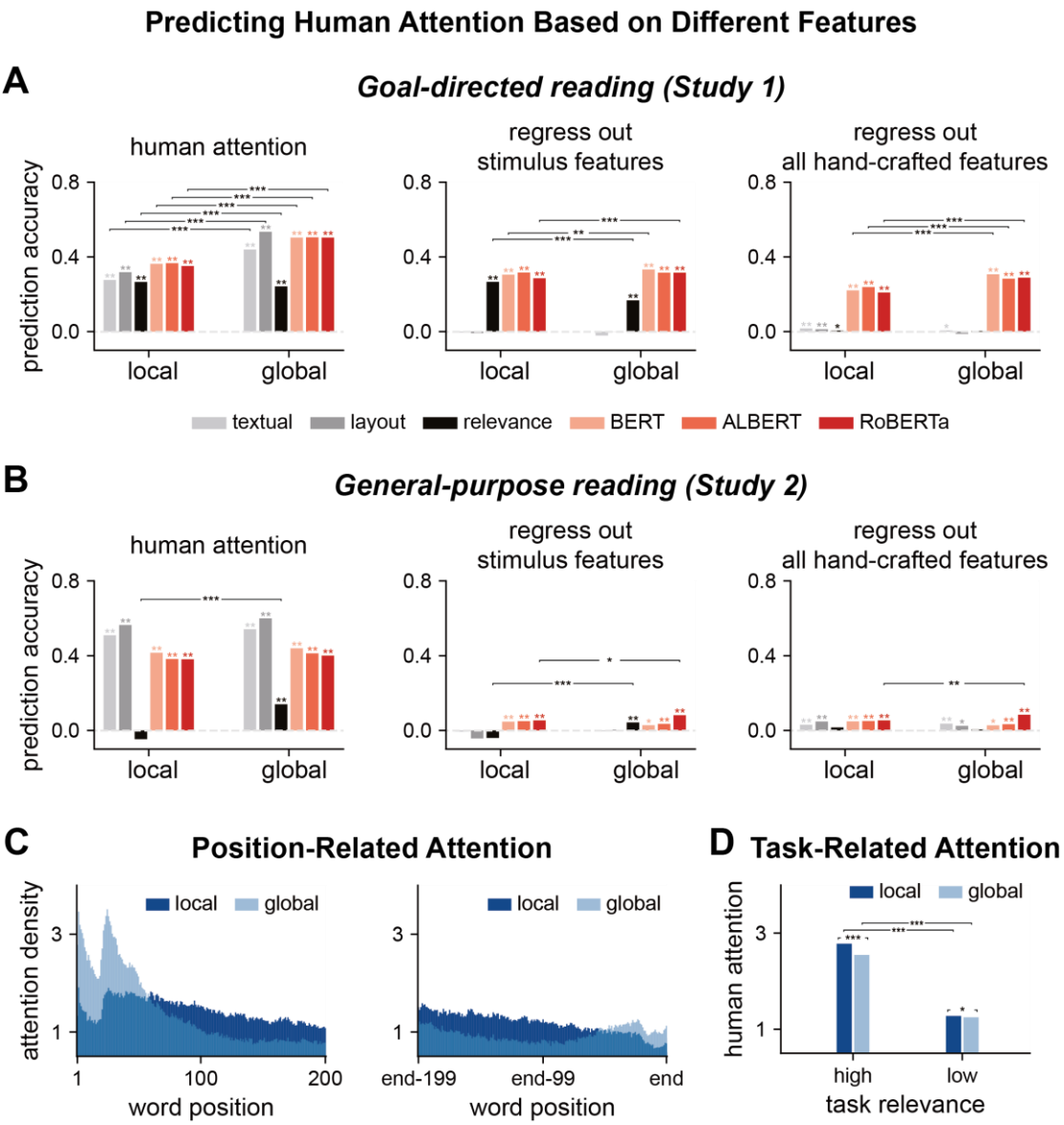

Fig. 3. Predicting human attention using different features. (**A** and **B**) Panels A and B show the results of Study 1 and Study 2, respectively. The left plots show how well different sets of features can predict human attention. In the middle and right plots, some features are regressed out from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

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 regression. The prediction accuracy, i.e., the correlation coefficient between the predicted and actual attention density, was significantly above chance ($P = 0.002$, permutation test, FDR corrected). Furthermore, the prediction accuracy was significantly higher for global questions than for local questions ($P = 1.4 \times 10^{-4}$, bootstrap, FDR corrected) (Fig. 3A, the left plot). We then used the same regression analysis to analyze whether the visual layout of a passage could also influence attention distribution. Here, layout features referred to features induced by line changes (see *Materials and Methods* for details), which could be processed without word recognition. The prediction accuracy for layout features was also statistically significant ($P = 0.002$, permutation test, FDR corrected).

Textual features and layout features characterized properties of the stimulus that were invariant across tasks. In the following, we investigated whether the task, i.e., to answer a specific question, also modulated human attention distribution. To characterize the top-down influence of task, we acquired annotations indicating each word's contribution to question answering, i.e., task relevance (see *Materials and Methods*). As shown in the left plot of Fig. 3A, we found that task relevance could indeed significantly predict human attention distribution ($P = 0.002$, permutation test, FDR corrected). Since task relevance was not a well-established modulator of reading attention, we further analyzed whether the task relevance effect could be explained by

197 the well-established textual and layout effects. In this analysis, we first regressed out
198 the influence of textual and layout features from the human attention distribution, and
199 found that the residual attention distribution could still be predicted by task relevance
200 (P = 0.003, permutation test, FDR corrected) (Fig. 3A, middle plot). These results
201 showed that the top-down reading goal, quantified by task relevance, could modulate
202 human attention, on top of lower-level stimulus features, i.e., textual and layout features.
203
204 The linear regression analyses revealed that textual features, layout features, and task
205 relevance all modulated human attention (see Fig. 2 for examples). The prediction
206 accuracy for different features ranged between 0.2 and 0.6, comparable to the prediction
207 accuracy of visual saliency models when predicting human attention to images[31,32].
208 Further analyses also revealed how these features modulated human attention. For
209 example, we found that participants generally attended more to the beginning of a
210 passage (Fig. 3C). Furthermore, this effect was stronger for global questions, which
211 potentially explained why stimulus features could better predict the attention
212 distribution for global questions. Additionally, it was also found that participants
213 attended more to words that are more relevant to the question answering task (Fig. 3D).
214
215 **Attention Distributions in Humans and DNN**
216 We then investigated whether DNN attention was comparable to human attention. The
217 general architecture of the models was illustrated in Fig. 1C. The input to the models
218 included all the words in the passage, integrated option, and 3 special tokens. One of

219 the special token, i.e., CLS, was the decision variable, based on the final representation
220 of which the DNN models decided whether an option was the correct answer or not. In
221 the following, we analyzed the attention weight between the CLS token and each word
222 in the passage (see *Materials and Methods* for details). In each layer of the DNN models,
223 the vectorial representation of the CLS token was updated by a weighted sum of the
224 vectorial representations of all input words and tokens. Therefore, the attention weight
225 on a word could reflect how heavily the word contributed to the decision variable, i.e.,
226 the CLS token.
227
228 We analyzed 3 DNN models, i.e., BERT[17], ALBERT[18], and RoBERTa[19], and the
229 question answering performance of the 3 DNN models was within the range of human
230 performance (Fig. 1B). Each of the 3 DNN models had 12 layers and each layer had 12
231 heads, each of which had a separate set of attention weights (Fig. 1CD). Consequently,
232 each word had 144 attention weights (12 layers × 12 heads). In the following, we first
233 tested whether the DNNs learned human-like attention distributions by attempting to
234 decode human attention distribution from the 144 DNN attention weights using linear
235 regression. Then, we analyzed whether the attention weights in different layers showed
236 different properties.
237
238 Although the DNN models were only trained to perform the reading comprehension
239 task and were blind to the human fixation data, it was found that the DNN attention
240 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 plot). The prediction accuracy of DNN attention weights was higher than that of textual features and task relevance. When 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 should be mentioned, however, 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 features 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 could 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 specific 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 ($P = 0.003$, 0.003, and 1 for textual features, layout features, and task relevance, permutation test, FDR corrected). For global questions, all three features could predict human attention distribution ($P = 0.003$, 0.003, and 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).

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.

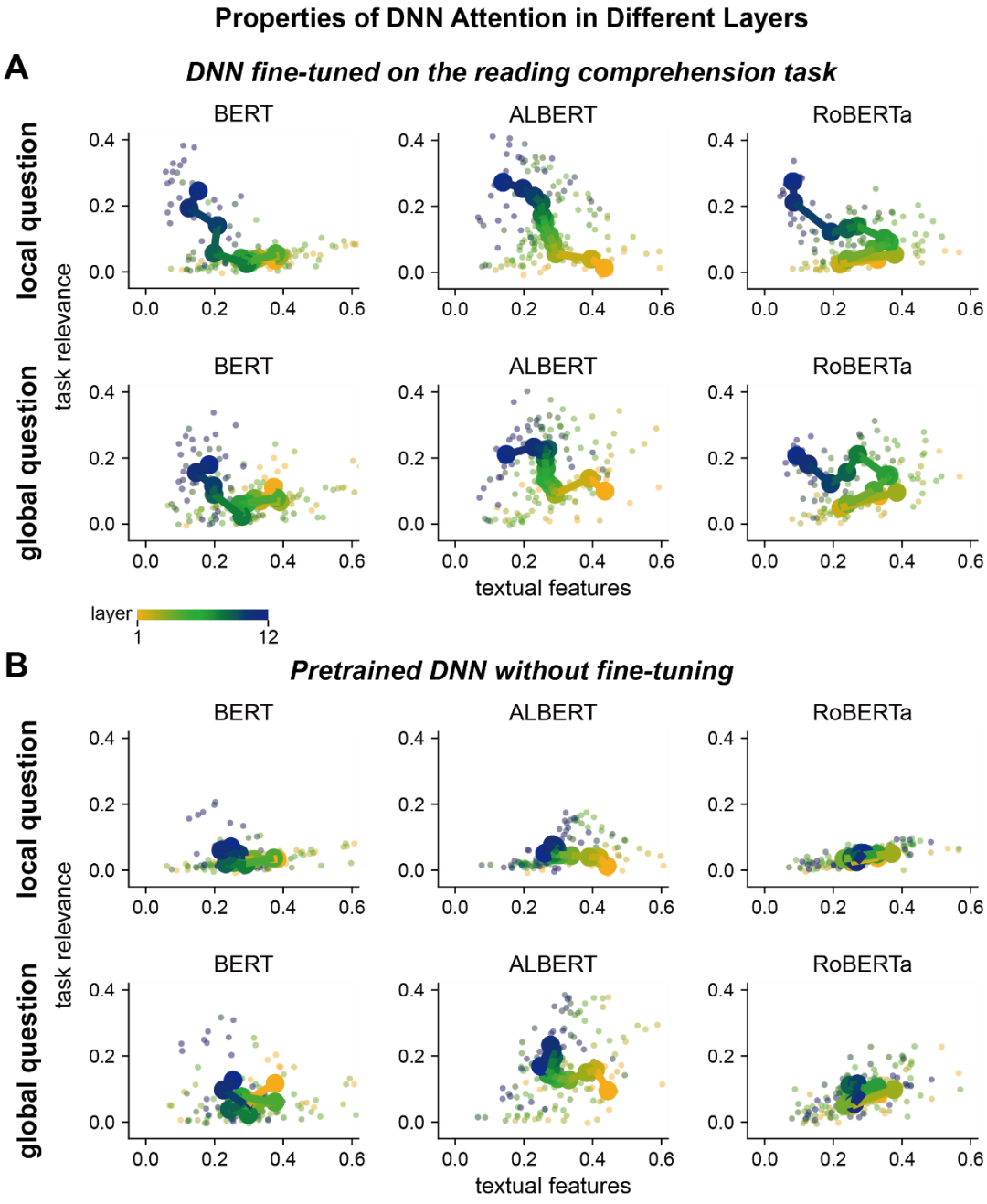

Fig. 4. Influence of stimulus features and top-down task on each DNN layer. The same regression analyses in Fig. 3 are employed to analyze how the DNN attention is affected by lower-level stimulus features and top-down task relevance. Panels A and B show the results for the DNNs fine-tuned based on the reading comprehension task and the pre-trained DNNs that receive no fine-tuning. 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 DNN are more sensitive to stimulus features. Deep layers of fine-tuned DNN, but not pre-trained DNN, are sensitive to task relevance.

DNN Attention in Different Layers

Previous studies have shown that different layers in DNN encoded different types of information[42-44]. In the following, we analyzed whether the properties of DNN attention weights differed across layers. Since human attention was influenced by both bottom-up stimulus features and top-down task goal, in the following we also analyzed how these features influenced the attention weights in each DNN layer. 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 (see *Materials and Methods*). Was the DNN attention mechanism mainly shaped by the pre-training process or the fine-tuning process? 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 DNN were sensitive to textual features in shallow layers but not sensitive to task relevance in deeper layers, suggesting that top-down attention in DNNs

emerged during fine-tuning using the reading comprehension task.

We then asked how the attention weights of DNN changed during fine-tuning and whether such changes were related to the performance of question answering. During fine-tuning, the structure of the DNN model remained but the parameters were adjusted. 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 performance (Fig. 5AD), especially for local questions (Fig. 5A). The trend is less clear for global questions and a potential explanation is that global questions concern the main topic of the passage and can be answered by paying attention to different sets of words. Deep layer's sensitivity to textual features, however, dropped during fine-tuning (Fig. 5BE). Therefore, fine-tuning directed deep layers' attention towards task relevant information, sacrificing the sensitivity to textual features. Additionally, we found that the similarity between DNN attention weights and human attention was also boosted by fine-tuning for local questions (Fig. 5C). This result further demonstrated that human-like attention in DNNs was the consequence of optimization of the reading comprehension task, instead of the consequence of more general pre-training language tasks.

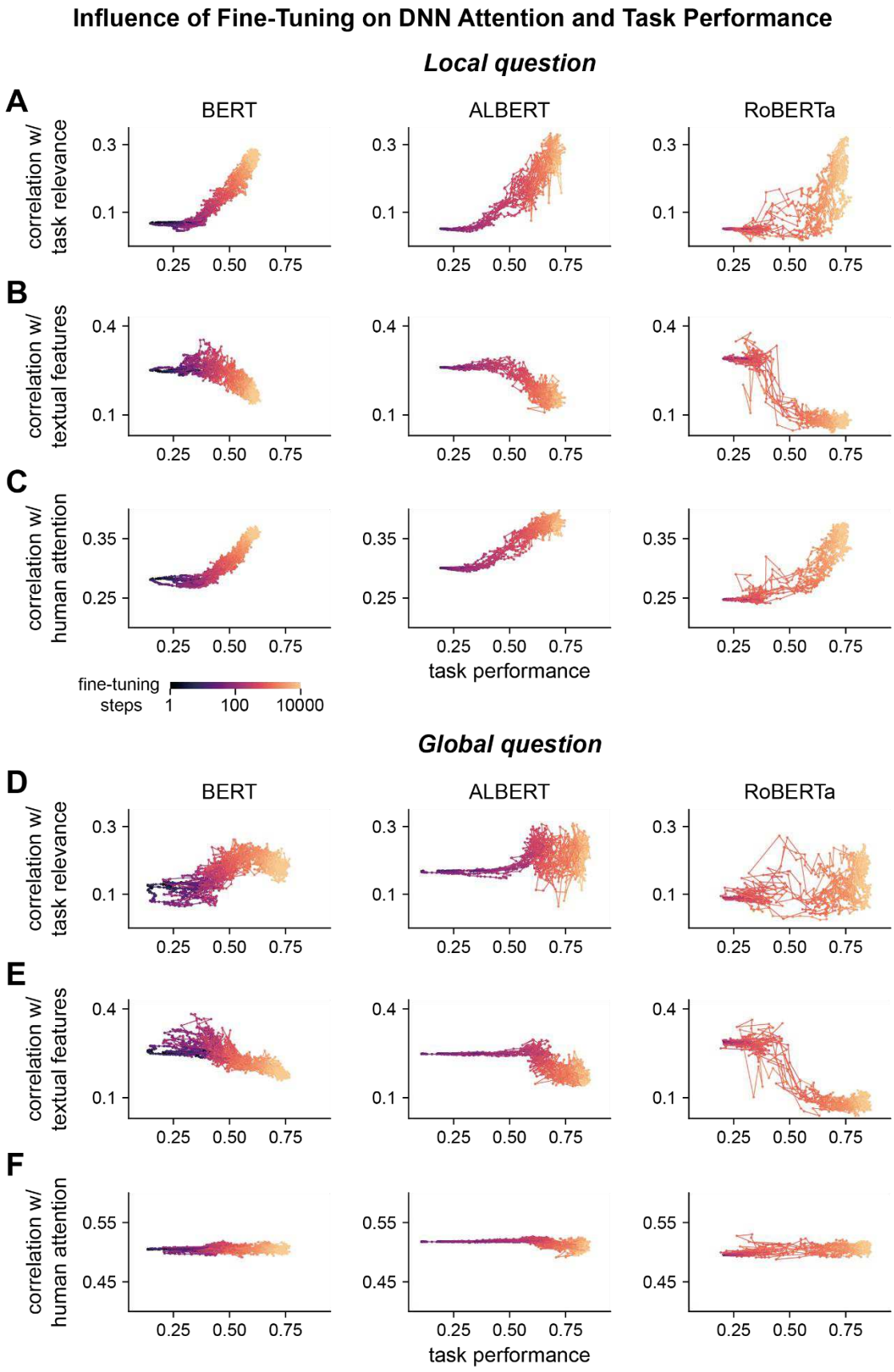

Fig. 5. Influence of fine-tuning on DNN attention and task performance. Each model is fine-tuned 10 times. Each data point denotes the result during a fine-tuning step (color coded), and steps from each run of fine-tuning was connected by a line.

(**A**, **B**, **D**, and **E**) The effect of fine-tuning on the attention weights in the last layer of DNN 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

Since attention is a key mechanism for both the biological brain and artificial neural networks, it provides a common ground to quantitatively compare biological neural computations and artificial neural computations. Such comparisons, however, are challenging since biological and artificial neural networks are investigated in different fields using very different approaches. The current study attempts to bridge this gap by building a large eye tracking dataset for a real-world reading task that is of interest to both the psychology and AI community. Based on these data, it is shown that, when optimized to perform a reading comprehension task, DNNs naturally evolve human-like attention distribution. On the one hand, the results indicate that the attention mechanism in DNNs could indeed be of biological relevance. On the other hand, it provides a plausible computational explanation for human attention distribution.

Computational models of biological attention

Deep neural network models of biological attention are best studied in vision. A large number of models are proposed to predict bottom-up visual saliency[31,32], and recently DNN models are also employed to model top-down visual attention. It is shown that, through either implicit[34,35] or explicit training[9,36], DNN can predict which parts of a picture relates to a verbal phrase, a task similar to goal-directed visual search[30]. The current study distinguishes from these studies in that the DNN model is not trained to predict human attention. Instead, the DNN models naturally generate human-like distribution when trained to perform the same task that humans perform. Therefore, the current study suggests that DNN models can potentially serve as a mechanistic, instead of a descriptive model of human attention during reading comprehension. Previous studies have also proposed mechanistic models of biological attention. It has been proposed that attention can be interpreted a mechanism to implement optimal decision making. For example, when faced with multiple conflicting cues, the brain can use attention to modulate, i.e., weight, the neural representation of each cue. It has been proposed that the brain attends to more informative and reliable cues to make an optimal decision[45-47]. The current results are generally consistent with this idea since both human and DNN attend to words that are relevant to task solving.

Attention during human reading

How human readers allocate attention during reading is an extensively studied topic. Eye tracking studies have shown that the readers fixate longer at, e.g., longer words,

words of lower-frequency, words that are less predictable based on the context, and words at the beginning of a line[48,49]. A number of models, e.g., the E-Z reader[37,50] and SWIFT[51], have been proposed to predict the eye movements during reading, either based on basic oculomotor properties or lexical processing[37]. These models can generate fine-grained predictions, e.g., which letter in a word will be fixated first. A limitation of these models, however, is that they are generally developed to explain the reading of simple sentences, instead of complex sentences or multi-line text. In contrast to these studies, the current study focuses on macroscopic distribution of attention, i.e., the distribution of total fixation time in the units of words, when readers read challenging multi-line text. Future studies can potentially integrate classic eye movement models with DNNs to explain the dynamic eye movement trajectory, possibly with a letter-based spatial resolution.

A more important difference between the current and previous studies on reading attention is that the current study investigates how attention is affected by a specific top-down reading goal. In previous studies, readers are generally instructed to read a sentence in a normal manner, not aimed to extract a specific kind of information. In the current study, however, readers know in advance what question they have to answer and this kind of reading can be viewed as a kind of information seeking behavior[52], and is also referred to as the reading-to-do task[38]. Previous studies have shown the reader's task may have heterogeneous influences on attention, depending on the task difficulty and skill level of readers[53,54]. Here, the task is demanding and the readers are highly

417 skilled to perform the task: The reading comprehension questions are selected from
418 exams and the time to answer each question is limited, leading to about 80% question
419 answering accuracy (Fig. 1). The participants are skilled since all Chinese students have
420 extensive practice in such reading comprehension questions in high school. Future work
421 is needed to quantify how the task and reading skills modulate human attention and
422 whether these effects can also be modeled by DNN models.
423
424 **Attention mechanisms in DNN**
425 In DNN NLP models, attention is a mechanism to selectively integrate information
426 across words (or other units of representations), and is typically implemented by
427 assigning different weights to different words. How the attention mechanism is
428 integrated with the neural network model and the role it plays, however, differ across
429 models[36,55]. In some models, attention is an explicit information integration mechanism.
430 For example, in many models[22,39,56-58], the representations of all words in a passage are
431 integrated with different weights to compute a representation of the passage. In these
432 models, ideally, words that contribute more to task solving should receive higher
433 weights.
434
435 In transformer-based models, the roles self-attention plays are highly diverse. Since
436 self-attention assigns a weight between every pair of inputs (including words and
437 special tokens such as CLS), it can capture a number of relationship between words,
438 e.g., co-reference and syntactic dependency[24-26]. 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. The current study demonstrated that 3 transformer-based models generate human-like attention through task optimization. It remains unclear whether other DNN models show similar properties. Nevertheless, the dataset and methods developed here can be easily applied to test whether other models also evolve human-like attention distribution, serving as a probe to test the biological plausibility of NLP models.

Interpretation of the attention mechanisms in DNN

Whether attention can increase the interpretability of DNN models is a topic that receives a considerable amount of debates. A number of studies have shown that the DNN attention weights are higher for words that are more important for the task[22,23,58,59]. Most of these studies, however, are based on visual inspection of a couple of examples. Other studies, however, find low correlation between attention weights and other measures of the importance of words, and therefore raise concern about whether the attention weights are interpretable[25,26,60,61]. The importance of a word, however, can be measured in many different ways, and no correlation with some importance measures does not indicate no contribution to task solving in other ways. Here, we show that, for 3 transformer-based models, the DNN attention weights correlate with human attention and are modulated by both textual features and the task. Furthermore, by analyzing the

fine tuning process (Fig. 5C), we found that DNN with human-like attention perform better at the task.

Here, it is revealed that different layers in DNN show different attention properties, with the deep layers being more sensitive to task relevance. This result is consistent with previous findings that artificial neurons in deep layers of DNN encode more abstract information, e.g., object information in convolutional networks[42] and syntactic information in BERT[43,44]. A recent study has also compared human eye movements and attention weights in the last layer of BERT, when participants evaluate whether a passage is an appropriate answer to a question. It is demonstrated that the human fixation time is more similar to the attention weights in BERT than the simple TF-IDF weights[27].

In sum, the current study demonstrates that, when DNN and humans perform the same reading comprehension task with comparable accuracy, the DNN attention weights resemble human attention measured by eye tracking. The results suggest that human attention distribution is shaped by the demand to optimally perform the task and the DNN attention can be interpreted as an approximation of human attention. The large set of eye tracking data in the current study can also motivate future computational modeling of human attention during natural reading tasks and be applied to test whether other NLP 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). All participants were native Chinese speakers and were college students or graduate students at 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[39]. We selected eight hundreds of high-school level questions from the test set of RACE and each question was associated with a distinct passage (117 to 456 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.,

[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.

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. We further acquired annotations about the relevance of each word to the question answering task. Details about the question types and the annotation procedures could be found in reference[62].

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, participants were given a familiarization session with 5 questions that were not used in the formal experiment. 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 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. After answering the question, they also rated their confidence about their answer on the scale of 1-4 (low to high). The confidence ratings were not analyzed.

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, and the main difference was 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. In Study 2, the correctness of the answer was also the prerequisite for bonus, and the amount of bonus decreased linearly with the duration of second-pass passage reading.

549

550 **Stimulus presentation and eye tracking**

551 The text was presented using the bold Courier New font, and each letter occupied 14 ×
552 27 pixels. We set the maximum number of letters on each line to 120 and used double
553 space. We separated paragraphs by indenting the first line of each new paragraph.
554 Participants sat about 880 mm from a monitor, at which each letter horizontally
555 subtended approximately 0.25 degrees of visual angle.

556

557 Eye tracking data were recorded from the left eye with 500-Hz sampling rate (Eyelink
558 Portable Duo, SR Research). The experiment stimuli were presented on a 24-inch
559 monitor (1920x1080 resolution; 60 Hz refresh rate) and administered using MATLAB
560 Psychtoobox[63]. Each experiment started with a 13-point calibration and validation of
561 eye tracker, and the validation error was required to be below 0.5º of visual angle.
562 Furthermore, before each trial, a 1-point validation was applied, and if the calibration
563 error was higher than 0.5º, a recalibration was carried out. Head movements were
564 minimized using a chin and forehead rest.

565

566 **DNN models**

567 We tested 3 popular transformer-based DNN models, i.e., BERT[17], ALBERT[18], and
568 RoBERTa[19]. ALBERT and RoBERTa were both adapted from BERT, and had the same
569 basic structure. RoBERTa differed from BERT in its pre-training procedure[19] while
570 ALBERT applied factorized embedding parameterization and cross-layer parameter

sharing to reduce memory consumption[18]. Following previous works[18,19], 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. 1C, for the i^{th} option, the input to DNN 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[18,19], 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 from data. The score was independently calculated for each option and then normalized using the following equation:

$$score_i = \frac{exp(s_i)}{\sum_{i=1}^{4} exp(s_i)}.$$

The answer to a question was determined as the option with highest score, and all the models were trained to maximize the logarithmic score of the correct option.

We fine-tuned DNN based on the training set of RACE. In the analyses shown in Figs. 5 and 6, the fine-tuning process was independently run 10 times. Each time, the training samples were fed in with a randomized order and nodes in the dropout layer were randomly eliminated. Results from the first run of fine-tuning was used for the main analysis reported in Figs. 2-4. All models were implemented based on HuggingFace[64] and all hyperparameters for fine-tuning were adopted from previous studies (Table S1). To isolate how the fine-tuning process modulated DNN attention, we also tested the pre-trained DNN that was not fine-tuned on RACE dataset, and compared it with the fine-tuned model (Fig. 4). Furthermore, we quantified how the properties of DNN attention changed throughout the fine-tuning process by analyzing models that received different steps of fine tuning. The steps we sampled were exponentially spaced between 1 and the maximum fine-tuning steps.

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[20]. 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. 1D). For each layer, the output of an attention head 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. 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 weight associated with the correct option.

Output of the attention module, i.e., C^h, was concatenated over all the 12 heads in each layer, and further processed by position-wise operations to generate the final representation of CLS in the layer[20]. Additionally, DNN used byte-pair tokenization which split some words into multiple tokens. We converted the token-level attention weights to word-level attention weights by summing the attention weights over tokens within a word[24,27].

636 **Human attention analysis and prediction**

637 We analyzed eye fixations during passage reading in Study 1 and the first-pass passage
638 reading in Study 2. For each word, the total fixation time was the sum of the duration
639 across all fixations that fell into the square area the word occupied. We averaged the
640 total fixation time across all participants who correctly answered the question, and
641 measured human attention using the attention density, i.e., the total fixation time
642 divided by the area a word occupied.

643

644 We employed linear regression to test whether a set of features could explain human
645 attention distribution. Four sets of features were analyzed, i.e., textual features, layout
646 features, task relevance, and DNN attention weights. The textual features included word
647 length, logarithmic word frequency estimated based on the British National Corpus [65],
648 ordinal position of a word in a sentence, ordinal position of a word in a passage, and
649 ordinal sentence number of a word. The layout features referred to the visual layout of
650 text, i.e., features induced by line changes, including the coordinate of the left most
651 pixel of a word, ordinal position of a word in a paragraph, ordinal row number of a
652 word in a paragraph, ordinal row number of a word in a passage. Task relevance was
653 annotated by humans, and the DNN attention weights included the 144 attention
654 weights from all layers and attention heads. In the regression analysis, human attention
655 density on word *w* was modeled using the following equation.

656

657 $$attention\_density_w = \sum_{j=1}^{J} \beta_j F_{w,j} + b + \varepsilon_w,$$

where F and ε denoted the features being considered and the residual error, respectively. The parameters β and b were fitted to minimize the mean square error. Each feature and the human attention distribution were normalized within a passage by taking the z-score. The prediction accuracy, i.e., the correlation between predicted attention and actual human attention, was calculated based on five-fold cross-validation. Each question type was separately modeled.

Statistical tests

We employed a one-sided permutation test to test whether the attention distribution predicted by a set of features significantly correlated with human attention. Five hundreds of chance-level prediction accuracy was calculated by predicting shuffled human attention. Specifically, the human attention density was shuffled across words and was predicted 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[66]. For example, to test whether the prediction accuracy differed between the 2 types of questions, all global questions were resampled with replacement 5000 times and each time the prediction accuracy was calculated based on the resampled questions, resulting in 5000 resampled prediction accuracy. If the prediction

accuracy for local questions was greater (or smaller) than N out of the 5000 resampled accuracy for global questions, the significance level of their difference was $2(N + 1)/5001$. When multiple comparisons were performed, the p-value was further adjusted using the false discovery rate (FDR) correction.

Acknowledgements

We thanks David Poeppel, Jonathan Simon, and Xunyi Pan for thoughtful comments on earlier versions of this manuscript; Yuran Zhang, Anqi Dai, Zhonghua Tang, and Yuhan Lu for assistance with experiments; Peiqing Jin and Cheng Luo for helpful discussions. 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 Nasr, K., Viswanathan, P. & Nieder, A. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances* **5**, eaav7903 (2019).

2 Kim, G., Jang, J., Baek, S., Song, M. & Paik, S.-B. Visual number sense in untrained deep neural networks. *Science Advances* **7**, eabd6127 (2021).

3 Sheahan, H., Luyckx, F., Nelli, S., Teupe, C. & Summerfield, C. Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron* **109**, 1214-1226. e1218 (2021).

4 Xu, Y. & Vaziri-Pashkam, M. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications* **12**, 1-16 (2021).

5 Ettinger, A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. In *Proc. Transactions of the Association for Computational Linguistics*, 34-48 (Association for Computational Linguistics, 2020).

6 Hale, J., Dyer, C., Kuncoro, A. & Brennan, J. Finding syntax in human encephalography with beam search. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics*, 2727-2736 (Association for Computational Linguistics, 2018).

7 Linzen, T., Dupoux, E. & Goldberg, Y. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. In *Proc. Transactions of the Association for Computational Linguistics*, 521-535 (Association for Computational Linguistics, 2016).

8 Judd, T., Durand, F. & Torralba, A. A benchmark of computational models of saliency to predict human fixations. In *Proc. 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 134-142 (SciTePress, 2016).

9 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).

10 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).

11 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).

12 Donhauser, P. W. & Baillet, S. Two distinct neural timescales for predictive speech processing. *Neuron* **105**, 385-393. e389 (2020).

13 Seidenberg, M. S. & McClelland, J. L. A distributed, developmental model of word recognition and naming. *Psychological review* **96**, 523-568 (1989).

14 Pinker, S. & Prince, A. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**, 73-193 (1988).

15 McClelland, J. L. & Elman, J. L. The TRACE model of speech perception. *Cognitive psychology* **18**, 1-86 (1986).

16 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).

17 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).

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

19 Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. Preprint

at https://arxiv.org/abs/1907.11692 (2019).

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

21 Brown, T. B. *et al.* Language models are few-shot learners. In *Proc. Advances in Neural Information Processing Systems*, 1877-1901 (Curran Associates, 2020).

22 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).

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

24 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).

25 Kovaleva, O., Romanov, A., Rogers, A. & Rumshisky, A. Revealing the Dark Secrets of BERT. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4365-4374 (Association for Computational Linguistics, 2019).

26 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).

27 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).

28 Corbetta, M. & Shulman, G. L. Control of goal-directed and stimulus-driven

attention in the brain. *Nat. Rev. Neurosci.* **3**, 201-215 (2002).

29 Petersen, S. E. & Posner, M. I. The attention system of the human brain: 20 years after. *Annu. Rev. Neurosci* **35**, 73-89 (2012).

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

31 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).

32 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).

33 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).

34 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 (CVPR)*, 6077-6086 (IEEE, 2018).

35 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).

36 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).

37 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).

38 Duffy, T. M. & Kabance, P. Testing a readable writing approach to text revision. *Journal of educational psychology* **74**, 733-748 (1982).

39 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).

40 Liu, S. S., Zhang, X., Zhang, S., Wang, H. & Zhang, W. M. Neural Machine Reading Comprehension: Methods and Trends. *Appl. Sci.-Basel* **9**, 3698-3742, doi:10.3390/app9183698 (2019).

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

42 Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *Proc. European conference on computer vision*, 818-833 (Springer, 2014).

43 Lin, Y., Tan, Y. C. & Frank, R. Open Sesame: Getting Inside BERT's Linguistic Knowledge. In *Proc. 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 241-253 (Association for Computational Linguistics, 2019).

44 Jawahar, G., Sagot, B. & Seddah, D. What does BERT learn about the structure of language? In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, 3651-3657 (Association for Computational Linguistics, 2019).

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

46 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).

47 Rao, R. P. Bayesian inference and attentional modulation in the visual cortex. *Neuroreport* **16**, 1843-1848 (2005).

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

49 Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **124**, 372-422 (1998).

50 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).

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

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

53 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).

54 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).

55 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).

56 Chen, D., Bolton, J. & Manning, C. D. A thorough examination of the cnn/daily mail reading comprehension task. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2358-2367 (Association for Computational Linguistics, 2016).

57 Zhang, Y., Marshall, I. & Wallace, B. C. Rationale-augmented convolutional neural networks for text classification. In *Proc. 2016 Conference on Empirical Methods in Natural Language Processing*, 795-804 (Association for Computational Linguistics, 2016).

58 Wang, Y., Huang, M. & Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In *Proc. 2016 conference on empirical methods in natural language processing*, 606-615 (Association for Computational Linguistics, 2016).

59 Zhu, H., Wei, F., Qin, B. & Liu, T. Hierarchical attention flow for multiple-choice reading comprehension. In *Proc. AAAI Conference on Artificial Intelligence*, 6077-6085 (AAAI, 2018).

60 Wiegreffe, S. & Pinter, Y. Attention is not not Explanation. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 11-20 (Association for Computational Linguistics, 2019).

61 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).

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

63 Brainard, D. H. The psychophysics toolbox. *Spatial vision* **10**, 433-436 (1997).

64 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).

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

66 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.

- SMNC0814.pdf