Cognitive Insights into Document Comprehension: The Role of Reading Order and Visual Attention in Human and Large Language Models

Qingxuan Wang[1], Hao Wang[✉1], Huiran Zhang[1], Chenhui Chu[2], Rui Wang[3] and **Pinpin Zhu[1]**
[1]School of Computer Engineering and Science, Shanghai University, Shanghai, China
[2]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[3]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Abstract

This study investigates how integrating human eye-tracking data into Large Language Models (LLMs) and Visual Large Language Models (VLLMs) can enhance document comprehension in tasks that require both linguistic understanding and visual attention, specifically Semantic Entity Recognition (SER) and Document Question Answering (DQA). Despite rapid advancements in AI-based document understanding, LLMs still face challenges in replicating the depth of human cognition, particularly in how reading order and visual attention affect comprehension. The results demonstrate that human reading order and the regions they focus on significantly impact performance in both tasks. Additionally, while LLMs do not need to fully mimic human reading sequences, their performance improves when their attention patterns align more closely with human visual strategies. This highlights the importance of incorporating cognitive-inspired attention mechanisms in AI systems, offering a path to better AI models that reflect human cognitive strategies in complex document understanding.

Keywords: Large Language Models; Cognitive Processing; Reading Order; Visual Attention

Introduction

The rapid advancement of document intelligence technologies has revolutionized how we access, process, and comprehend information. Visually Rich Documents (VRDs)—comprising tables, graphs, charts, and other visual elements—are particularly pervasive due to their dense informational content. However, understanding these documents requires more than just processing textual data. It necessitates the integration of visual elements, layout structures, and cognitive mechanisms that align with human perceptual and cognitive processes (Treisman & Gelade, 1980; Heaton & Hummel, 2019; H. Wang et al., 2023; Penzkofer, Shi, & Bulling, 2024; Ding, Lee, & Han, 2024).

Human cognition plays a critical role in interpreting VRDs. Intuitively, figure 1 illustrates how human readers employ advanced cognitive strategies, such as reading order (Rayner, 1998; Bammel & de Oliveira, 2023) and visual attention (Dambacher, Kliegl, Hofmann, & Jacobs, 2006; Sood et al., 2023), to navigate complex document layouts. These mechanisms, shaped by our long-term evolution, enable individuals to efficiently prioritize relevant information and seamlessly integrate textual, visual, and layout data during reading. In contrast, despite the significant strides made by Large Language Models (LLMs) like ChatGPT (OpenAI, 2023) and LLaMA (Touvron et al., 2023; Dubey et al., 2024), as well as Visual Large Language Models (VLLMs) (Radford et al., 2021; GLM et al., 2024; P. Wang et al., 2024), in text comprehension, they still struggle to replicate the depth of human cognitive abilities when processing VRDs.

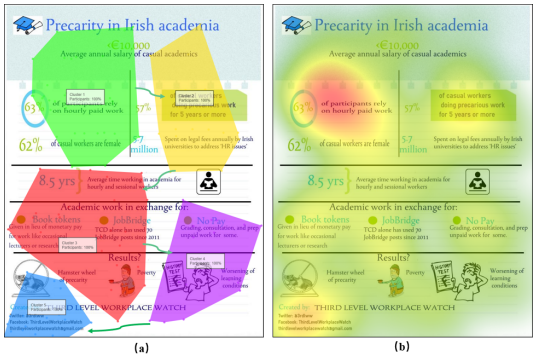

Figure 1: Visualization of human behavior in reading a VRD: (a) reading order and (b) visual heatmap.

This raises an important question: **Should AI models mimic human cognitive behavior to enhance their understanding of VRDs?** Previous studies like LayoutReader (Z. Wang, Xu, Cui, Shang, & Wei, 2021) and DocTrack (H. Wang et al., 2023) provide valuable insights into how human reading sequences benefit document AI models, but they capture only a small fraction of the broader cognitive picture. While human reading is not simply about following a linear order, it is driven by ingrained attentional mechanisms—evolutionary adaptations that allow rapid scanning and visual focus on critical document areas (Vaswani et al., 2017; Guo et al., 2021). These processes guide efficient document comprehension by integrating textual, visual, and layout information through a combination of visual and cognitive cues.

To address this gap, we argue that current datasets fail to accurately capture human cognitive processes involved in VRD comprehension (Reichle, Rayner, & Pollatsek, 2003; Holsanova, Rahm, & Holmqvist, 2006; Mathews, Xie, & He, 2016). Eye-tracking data, which offers valuable insights into how humans integrate visual elements and layout information, is crucial in this context. To bridge this gap, we introduce a novel dataset, INFOGRAPH-146, collected via crowd-

[1]Corresponding author: wang-hao@shu.edu.cn.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

sourcing, comprising detailed eye-tracking data from 36 participants. This dataset captures a broader spectrum of cognitive patterns involved in VRD comprehension, focusing not only on reading order but also on regions of interest (ROIs) where human gaze is concentrated. By generating attention heatmaps (Fig. 1b), we provide a comprehensive representation of human cognitive strategies, essential for enhancing both human-centered and machine understanding of VRDs. Building on this dataset, we propose a series of experiments that integrate cognitive science insights with AI techniques, exploring how machine models can better align with human cognitive processes.

Contributions: Our findings make several important contributions to the field of AI-driven VRD understanding:

- We show that the refined `Zrule` reading order, which is favored by LLMs and VLLMs, is more suitable for machine models than the traditional human reading order.
- We propose a method for injecting human visual attention into LLMs and VLLMs, enabling these models to more accurately replicate human reading behavior.
- We demonstrate that aligning AI models with human cognitive patterns—such as reading order and attentional regions—can significantly enhance the performance of LLMs and VLLMs on VRD comprehension tasks. This highlights the potential of cognitive-inspired attention mechanisms in improving AI comprehension of VRDs, emphasizing the importance of harmonizing machine models with human cognitive processes for more effective document understanding.

Related Works

Eye-Tracking

Eye-tracking has become a fundamental tool for understanding the cognitive processes involved in human reading (Rayner, 1998). By capturing detailed real-time data on eye position, movement, and pupil size, eye-tracking systems provide valuable insights into how individuals process and navigate both textual and visual information. These systems identify areas of interest by tracking eye movements, enabling researchers to analyze how attention is allocated across different regions of a document. Eye-tracking technology has been applied across a wide range of disciplines. In visual systems research, it helps investigate how the brain processes visual stimuli (Barr, 2008). In psychology and neuroscience, it is used to explore the neural mechanisms behind attention and perception (Migliaccio, MacDougall, Minor, & Della Santina, 2005). In psycholinguistics and healthcare, eye-tracking has been employed to study reading behaviors and cognitive load (Park, Subramaniyam, Hong, Kim, & Yu, 2017). It also plays a key role in user experience and interaction studies, where it informs design and usability improvements (Lukander, 2016). Eye-tracking has proven valuable in professional performance and consumer research (Hang, Yi, & Xianglan, 2018), clinical research and education (Colliot & Jamet, 2018), and transportation (Noland, Weiner, Gao, Cook, & Nelessen, 2017), providing data on attention and behavior patterns. In addition to these applications, eye-tracking is increasingly used in sports performance analysis (Obaidellah, Al Haek, & Cheng, 2018), product design (Sharafi, 2015), virtual reality (Clay, Konig, & Konig, 2019), and assistive technologies for the elderly and people with special needs, helping to enhance user interaction and overall experience. Eye-tracking also advances AI tasks (Barrett, S., K., & T., 2016; Y. Zhang et al., 2024; Yan, Zhang, & Zhang, 2024; Hollenstein & K., 2019; Y. Zhang et al., 2024). Through its diverse applications, eye-tracking continues to deepen our understanding of human cognition and offers practical benefits across various fields.

Reading Order

Human reading order refers to the sequence in which people scan text, typically left-to-right or top-to-bottom in languages like English and Chinese, while languages like Japanese and Arabic follow right-to-left or vertical conventions. Eye-tracking studies show that readers naturally follow these patterns (J. Henderson & Ferreira, 1993), with their movements influenced by factors such as text difficulty, reader expertise, and content nature (Reichle et al., 2003; Schilling, Rayner, Chace, Swaab, & Liversedge, 2006; Clifton et al., 2016). However, VRDs feature complex layouts combining text, images, and other visual elements, prompting readers to adapt their scanning strategies based on the document's design (Heard, Rakow, & Foulsham, 2017). Typically, readers start by surveying the layout, focusing on key elements like headings or figures, before delving into more detailed sections. This context-driven approach helps readers connect the textual and visual components effectively.

Visual Attention

Attention, a key concept in psychology, neuroscience, and AI, governs how cognitive resources are allocated to process relevant information. Classic models, such as the limited resource theory and bottleneck model (Broadbent, 1958), emphasize attention's role in prioritizing content to prevent overload. In reading, attention shifts based on content and context (J. M. Henderson, 2003), with multitasking negatively impacting performance (Pashler, 1994). Recent AI models for VRD understanding leverage eye-tracking data and visual search theory to mimic human attention, focusing on key regions like headings and figures (Rayner, 1998). By integrating cognitive models like ACT-R (Anderson, 2005) and EMMA (Salvucci, 2001), AI systems can better prioritize relevant content and simulate human attention shifts in complex document layouts.

Human Data Collection

Participants

We recruited 36 participants (27 males, 9 females, average age 22), all undergraduate or graduate students majoring in

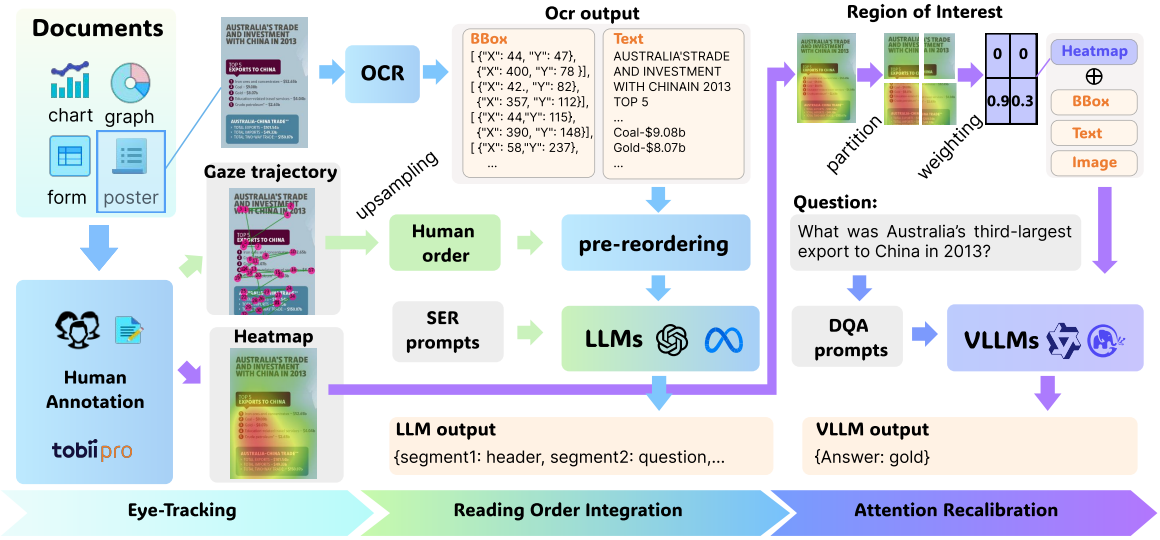

Figure 2: Integration pipeline of reading order and visual heatmap into LLMs/VLLMs: While we demonstrate the incorporation of the visual heatmap into VLLMs, it is possible to integrate it into LLMs by combining these weights with bounding boxes for input.

computer science or artificial intelligence. They were ESL learners with fluency levels equivalent to an IELTS score of 6.0, TOEFL score of 80, or CET-6. Participants underwent training and signed the research ethics statement. Their responses were checked for accuracy, with incorrect answers discarded and reconducted when needed to ensure full correctness. Participants were grouped into sets of three, with each group testing 50 documents. This grouping and collection method ensured diverse reading strategies, providing a rich dataset for analysis. Finally, each document had at least three human annotations.

Eye-tracking Setup and Documents

To collect eye-tracking data, we used a Tobii Pro TX300 eye tracker and Tobii Studio software. Participants read 150 documents selected from the INFOGRAPHICVQA (Mathew et al., 2021) dataset on a 24-inch 1080P HP monitor. These documents covered a variety of formats, from simple forms to complex, visually-rich layouts with both text and graphics. After reviewing the documents by experts, some were discarded due to the low quality of annotations, resulting in the final set of 146 documents, namely INFOGRAPH-146. This diverse selection allowed for a thorough analysis of how participants integrate textual and visual information during document comprehension.

Experiment Procedure

Eye movements, including gaze points, fixation duration, frequency, saccade distance, and pupil size, were recorded during the reading process. Data anomalies were corrected, addressing missing or abnormal points. Assessing the quality and representativeness of the data by considering factors such as consistency, correlation, and task performance, we ended up manually selecting the best annotation to use as human reference data from three human annotations.

Experimental Setup

LLMs and VLLMs

We evaluated several LLMs, including LLaMA3-8B, GPT3.5, and GPT4.0, alongside VLLMs such as GLM-4V-9B and Qwen2-VL-7B.

Task Definition

We evaluted two main subtasks of VRD understanding:

- **SER Task**: This task focuses on identifying and classifying OCR segments within documents into four categories such as `header`, `question`, `answer`, and `other`.
- **DQA Task**: The goal of this task is to generate accurate answers to user's question based on the content of VRDs.

For evaluation, we measure precision, recall and F1 scores for the SER task and F1 and ANLS (Average Normalized Levenshtein Similarity) for the DQA task.

Datasets

We examine the impact of reading order as well as visual attention using the INFOGRAPH-146 dataset on the DQA task. To ensure comparability with other works, we also compare the performance of the methods on public SER datasets: FUNSD (Jaume, Ekenel, & Thiran, 2019), SEAB (J. Zhang, Wang, & Luo, 2022), and SORIE (Huang et al., 2021).

SER Prompt

The model is prompted to classify segments output by the previous OCR(Girshick, 2015; Smith, 2007) engine based on both its content and spatial coordinates, without requiring additional annotations. To ensure the models understand the task, we provided basic instructions explaining that

the goal is to label each text segment with an entity category. Additionally, we included a simple, unrelated example to clarify and control the input-output format. For example, in a question-answer pair "`Port Of Loading:`" and "`San Francisco, California`", the former is classified as a `question`, and the latter is labeled as `answer`. This allows for consistent output, as illustrated in the following format: {"`Port Of Loading:`": "`question`", "`San Francisco, California`": "`answer`" }, which helps the model understand the required format for both input and output, ensuring consistency across classifications.

DQA Prompt

For the Document Question Answering (DQA) task, the model receives a user query and generates a response based on a structured prompt. This ensures that the model's answer is relevant to the query. For example, if asked, "`What is the official language of Spain?`", the system would generate: {"`answer`": "`Spanish`"}. If no relevant information is found, the response will be: {"`answer`": "`Not found`"}. This approach effectively integrates document content with user queries, improving the accuracy of the answers and the model's performance on tasks involving the comprehension of VRDs. Note that the example used as a format never appeared in the test dataset.

Reading Order Integration

Pipeline

To better align with human reading behaivor, we employ a pre-reordering technique (H. Wang et al., 2023) in the preprocessing stage of VRD analysis. The core idea of pre-reordering is to rearrange the document's content in a way that approximates the natural reading order of humans, which is inspired by the field of statistical machine translation (Neubig, Dyer, Yoshino, & Matsumoto, 2012; Nakagawa, 2015), where pre-reordering is used as a preprocessing step to reorder the input text and improve translation quality by optimizing sentence structures. In the context of our study, prereordering is not simply a technical procedure; it serves a more profound purpose: it allows us to mimic the rough trajectory of human reading documents and align this process with multi-modal input features including text, bounding box (BBox), and visual information. The middle part of Figure 2 illustrates the details of processing pipeline for reading order integration. After prereordering the OCR outputs, we input text and bounding box information formatted in a JSON file into LLMs. This helps us to understand how different reading sequences affect the effectiveness of document comprehension tasks.

Order Generation

1) `XYcut`: Gu et al. (2022) uses the XYCut algorithm to segment the document based on horizontal and vertical projection profiles, then locally determines the reading order of document elements. This approach improves the model's ability to understand complex document layouts, enhancing performance in VRD understanding tasks like entity recognition and relation extraction.

Algorithm 1: Z-Rule Pre-reordering

Input: b: A list of OCR bounding boxes.
Output: Resorted list according to bounding box.

Function `generate_z_rule_order(`b`)`**:**
  Sort b by y_{tl}, then by x_{tl};
  for $i \leftarrow 0$ **to** $len(b) - 1$ **do**
    for $j \leftarrow i$ **to** 0 **do**
      if $abs(b[j+1].y_{tl} - b[j].y_{tl}) < thresholds$ ***and*** $b[j+1].x_{tl} < b[j].x_{tl}$ **then**
        Swap $b[j]$ and $b[j+1]$;
      else
        break;
  return b;

2) `human-order`: As mentioned earlier, this method directly utilizes the natural sequence of human reading captured through eye-tracking. The document input sequence for the LLM is rearranged through upsampling based on the human eye movement reading order, fully reflecting the natural order of human reading. This reveals how human readers prioritize different sections of a document. The effect of using human order in the VRD understanding task directly reflects the effectiveness of human reading order in large models.

3) `agent-order`: Using a LayoutLMv3-based agent to learn human reading order (H. Wang et al., 2023), this method generates human-like reading sequences based on multimodal document features. By training with a large amount of real human reading paths as reference data, the model learns to prioritize elements typically focused on by humans, such as titles and keywords, to adjust the order. It uses a contrastive loss function to distinguish between correct and incorrect sequences. Generally, the generated order closely resembles the human-provided order.

4) `Zrule-order`: Given the input consists of a list of bounding boxes, each defined by the top-left coordinates $(x_{\text{tl}}, y_{\text{tl}})$ and the bottom-right coordinates $(x_{\text{br}}, y_{\text{br}})$, we design algorithm proceeds as shown in Algorithm 1. `Zrule-order` ensures the bounding boxes are ordered first by their vertical position and then by their horizontal position within each row. This addresses the issue of inconsistent reading order caused by OCR errors. The algorithm adjusts the order if two bounding boxes are close enough in vertical distance (within a specified threshold) and the horizontal position of the subsequent box is smaller, maintaining a correct left-to-right, top-to-bottom reading flow. To our surprise, we found this simple human-like order is more suitable for LLMs and often outperforms other orders.

Model	Order	SER									DQA	
		FUNSD			SEAB			SORIE			INFOGRAPH-146	
		P	R	F1	P	R	F1	P	R	F1	F1	ANLS
LLaMA3-8B	default	53.00	29.44	37.85	36.43	77.86	49.64	51.39	63.44	56.79	18.51	18.51
	XYcut	76.42	42.85	54.91	37.83	78.91	51.14	**52.21**	72.10	60.57	20.16	31.16
	Human-order	74.63	36.02	48.59	39.09	**87.10**	53.96	-	-	-	19.49	26.07
	Agent-order	79.72	44.53	57.14	**40.30**	83.30	**54.32**	50.21	**77.59**	60.96	20.78	26.78
	Zrule-order	**81.61**	**59.43**	**68.77**	37.57	79.89	51.46	50.98	77.07	**61.36**	**22.65**	**31.67**
GPT3.5	default	69.93	41.83	52.34	31.21	58.04	40.59	67.88	56.02	61.38	22.01	26.45
	XYcut	70.43	53.28	60.67	31.45	**61.10**	41.52	66.27	68.07	67.16	27.35	30.76
	Human-order	78.28	53.44	63.52	31.04	56.17	39.98	-	-	-	25.49	31.51
	Agent-order	76.58	62.33	68.72	31.28	59.56	41.02	71.91	**76.52**	74.15	24.68	31.57
	Zrule-order	**78.88**	**66.59**	**72.21**	**31.82**	58.81	**41.53**	**76.19**	72.72	**74.52**	**28.43**	**33.33**
GPT4.0	default	83.45	61.67	70.93	39.45	45.38	42.76	57.00	**87.04**	68.89	31.08	38.76
	XYcut	84.29	65.02	73.41	**50.88**	63.20	56.38	60.67	84.37	70.58	35.01	41.76
	Human-order	85.86	63.63	73.09	47.49	59.96	53.00	-	-	-	36.22	43.55
	Agent-order	**87.74**	74.06	80.64	46.50	59.06	52.03	74.45	85.62	79.65	35.71	44.19
	Zrule-order	87.11	**75.31**	**80.78**	50.47	**67.62**	**57.81**	**79.16**	83.12	**81.09**	**37.49**	**44.53**

Table 1: Impact of reading order integration on SER and DQA tasks across different LLMs. **Bold** indicates the best result in each group, while underlined indicates the second-best result. `Zrule-order` achieved the best results in 11 out of 12 groups.

Results and Analysis

Table 1 presents the impact of various reading order integration methods—namely, `default`, `XYcut`, `Human-order`, `Zrule-order`, and `Agent-order`—on the SER and DQA tasks across three different LLMs: LLaMA3-8B, GPT3.5, and GPT4.0. The key results highlight the significant improvements that certain reading order methods bring, especially in the context of entity recognition and question answering. The integration of specific reading order methods significantly improves performance in both SER and DQA tasks. While `Zrule-order` generally yields the best results across models, `Agent-order` offers strong performance with notable second-best results, particularly in tasks with multimodal content. This suggests that learning-based methods like `Agent-order` provide a more flexible and dynamic approach to generating reading orders, especially when the document layout and content vary in complexity. There are some key findings. `Zrule-order` consistently outperforms all other methods across all models, particularly excelling in F1 scores for the SER tasks. This suggests that this rule-based reading order method better mimic human reading patterns, leading to superior entity recognition performance across varying datasets. `Agent-order` performs well, particularly in tasks with multimodal content, like SEAB and INFOGRAPH-146. It consistently achieves second-best results, especially for tasks where understanding the context or spatial layout of the document is crucial. `Human-order` performs well for LLaMA3-8B and GPT3.5, suggesting that human-like reading order can provide substantial improvements in recognizing entities. However, this method does not scale as well with larger models (GPT4.0), where rule-based methods like `Zrule-order` are more effective. `XYcut` performs better than the default order but generally lags behind `Agent-order` and `Zrule-order`. This method is effective but may not capture the full complexity of multimodal and spatial relationships as well as the other approaches.

Attention Recalibration

Pipeline

1) Heatmap Injection: The eye-tracking device records fixation points and gaze duration, which are aggregated into a heatmap, where the intensity of each pixel reflects the fixation frequency or duration. For VLLMs, the heatmap is directly input into the visual encoder. For LLMs, we convert it into a heat matrix, aligning it with bounding box coordinates and extending OCR bounding box entries with heat values in the JSON file (see Figure 2, right). The heat matrix is normalized to the $[0, 1]$ range for consistency across documents. A prompt is added to help both LLMs and VLLMs interpret the data, highlighting regions with the most attention, thus enhancing the integration of visual and textual information.

2) Dataset Splitting Using Saliency Scores: In DQA task, human readers typically allocate more attention to regions where the answer is likely to appear. The more prominent a region (indicated by red areas in the heatmap), the more cognitive effort humans invest in searching for the answer. To examine the impact of ROI on the DQA task, we computed the average saliency scores at three different grid granularities: 3×3, 5×5, and 7×7 regions. For each region, we first calculated the color difference between the heatmap and the original image in the HSV color space. Based on this color difference, we assigned a saliency score to each region, where green, yellow, and red regions correspond to progressively higher saliency scores (green: 40-60, yellow: 60-80, red: 80-100). Regions with a score less than 40 were considered less attention loads. Using the overall average saliency score for each document, we then divide the INFOGRAPH-146 dataset into three subsets: low, medium, and high, respectively.

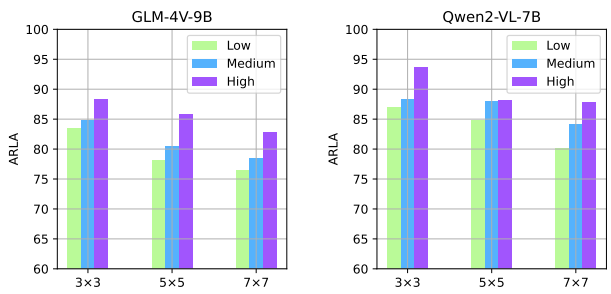

Figure 3: Answer Region Localization Accuracy (ARLA) of different sliancy sub-datasets (Low, Medium and High) with different grid granularities.

Model	Dataset	Saliency	ARLA	F1	ANLS
GLM-4V 9B	Low	73.06	79.33	41.48	59.66
	Medium	84.61	81.20	41.89	60.50
	High	**91.29**	**85.62**	**43.01**	**62.32**
Qwen2-VL 7B	Low	73.06	83.94	39.77	62.82
	Medium	84.61	86.82	40.16	63.05
	High	**91.29**	**89.90**	**42.59**	**63.94**

Table 2: Comparison of VLLMs' performance across datasets with saliency and ARLA scores.

Results and Analysis

Whether the visual attention of LLMs is consistent with that of humans? We further evaluated the performance of the VLLMs on each subset using a specific experimental design: a carefully crafted prompt allowed the VLLMs to identify the region most likely to contain the answer, resulting in Answer Region Localization Accuracy (ARLA), based on the predicted region. As shown in Figure 3, both GLM-4V-9B and Qwen2-VL-7B exhibited best performance in higher saliency regions (High). Performance improved as the saliency of the regions increased, with the best results observed in the 7×7 grid (higher granularity). Overall, Qwen2-VL-7B performed better across all conditions, higher saliency regions enhance model performance in DQA tasks, and ROI segmentation based on saliency could further improve performance.

Whether human visual attention has a positive impact on the VQA task for VLLMs? Table 2 shows that both GLM-4V-9B and Qwen2-VL-7B models perform better as the focus shifts to higher saliency regions, with the highest scores in the High saliency subset. As saliency increases, both ARLA (Answer Region Localization Accuracy) and performance metrics (F1, ANLS) improve, indicating that the models are more accurate in locating answers in regions with higher human attention. Qwen2-VL-7B outperforms GLM-4V-9B in ARLA, F1, and ANLS in the Medium and High subsets, suggesting it better localizes answers. Overall, the results confirm that leveraging human attention patterns through ROI segmentation significantly enhances model performance in DQA tasks.

	Modalities		Order	HeatMap	F1	ANLS
	BBox+Text	Image	+Zrule			
LLaMA3 8B	✓				18.51	18.51
	✓			✓	**20.64**↑	**23.91**↑
	✓		✓		22.65	31.67
	✓		✓	✓	**22.78**↑	**32.45**↑
GPT3.5	✓				22.01	26.45
	✓			✓	**23.17**↑	**28.89**↑
	✓		✓		28.43	33.33
	✓		✓	✓	**27.19**↓	**33.82**↑
GPT4.0	✓				31.08	38.76
	✓			✓	**32.67**↑	**40.03**↑
	✓		✓		37.49	44.53
	✓		✓	✓	**37.66**↑	**45.21**↑
GLM-4V 9B		✓			42.18	61.70
		✓		✓	**43.32** ↑	**64.42** ↑
	✓	✓			43.52	61.30
	✓	✓		✓	44.51	65.41
	✓	✓	✓		43.10	65.04
	✓	✓	✓	✓	**47.25** ↑	**66.30** ↑
Qwen2-VL 7B		✓			41.06	63.74
		✓		✓	**43.53** ↑	**65.38** ↑
	✓	✓			40.95	62.81
	✓	✓		✓	42.23	65.91
	✓	✓	✓		42.50	66.29
	✓	✓	✓	✓	**44.73** ↑	**67.39** ↑

Table 3: Exploring the impact of synergizing reading order and attention recalibration on DQA Tasks.

How Human Cognition Enhances LLMs

Table 3 presents the results of different LLMs and VLLMs in the DQA tasks with varying modalities and configurations, focusing on the synergy of reading order and attention recalibration using different strategies. The introduction of attention recalibration and the `Zrule-order` generally improves the performance of most models in the DQA tasks, with the best results typically seen when both methods are applied together. Among the models, Qwen2-VL-7B and GPT4.0 show the most consistent and significant gains. Specifically, Qwen2-VL-7B achieves the highest scores in F1 and ANLS (F1 = 44.73, ANLS = 67.39), while GPT4.0 also shows notable improvements, particularly when all techniques are applied. While LLaMA3-8B and GPT3.5 show improvements with recalibration and Zrule, the gains are less pronounced. These results suggest that models like Qwen2-VL-7B and GPT4.0 are particularly well-suited for leveraging attention recalibration and order adjustments, enhancing their performance in visual question answering tasks.

Conclusion

While fully replicating human comprehension of VRDs remains a challenge, our findings suggest that human cognitive processes, particularly reading order and attention, can positively influence the performance of LLMs/VLLMs. This highlights the potential benefits of integrating human-like attention patterns into AI systems. These insights offer valuable guidance for developing more effective future AI models.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Young Program: 62306173) and JSPS KAKENHI Program (JP23H03454).

References

Anderson, J. R. (2005). *Cognitive psychology and its implications*. Macmillan.

Bammel, M., & de Oliveira, G. S. (2023). Reading comprehension as embodied action: Exploratory findings on nonlinear eye movement dynamics and comprehension of scientific texts. In *Annual Meeting of the Cognitive Science Society* (pp. 2333–2340).

Barr, M. (2008). Visual systems and eye-tracking technology: An overview. *Journal of Visual Systems*, *12*, 45-59. doi: 10.1016/j.jvis.2008.01.005

Barrett, D., S., A. B. K. L., K., A. T., & T., M. (2016). Using eye-tracking data for part-of-speech tagging. In *Proceedings of the Annual Conference on Computational Linguistics (ACL).*

Broadbent, D. E. (1958). *Perception and communication*. Pergamon Press. doi: 10.1037/10037-000

Clay, G., Konig, S., & Konig, P. (2019). Virtual reality and eye-tracking: Advances and applications. *Journal of Virtual Reality and Broadcasting*, *16*, 20-30. doi: 10.3233/JVR-180986

Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith rayner's 40year legacy. *Journal of Memory and Language*, *86*, 1-19.

Colliot, E., & Jamet, E. (2018). The application of eye-tracking in clinical and educational settings. *Journal of Clinical Psychology*, *44*, 112-118. doi: 10.1002/jclp.22541

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89-103.

Ding, Y., Lee, J., & Han, S. C. (2024). *Deep learning-based visually rich document content understanding: A survey.*

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... et al. (2024). The Llama 3 Herd of Models. *CoRR*, *abs/2407.21783*. doi: 10.48550/ARXIV.2407.21783

Girshick, R. (2015). Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)* (p. 1440-1448). doi: 10.1109/ICCV.2015.169

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., ... Wang, Z. (2024). *ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools.*

Gu, Z., Meng, C., Wang, K., Lan, J., Wang, W., Gu, M., & Zhang, L. (2022). Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 4573-4582). doi: 10.1109/CVPR52688.2022.00454

Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T., ... Hu, S. (2021). Attention mechanisms in computer vision: A survey. *CoRR*, *abs/2111.07624*.

Hang, Y., Yi, F., & Xianglan, Z. (2018). Consumer behavior analysis using eye-tracking in marketing research. *Journal of Marketing Research*, *57*(4), 635-644. doi: 10.1509/jmr.16.0495

Heard, C. L., Rakow, T., & Foulsham, T. (2017). The role of presentation order and orientation on information search and evaluations: An eye-tracking study. *Cognitive Science*, 2174-2179.

Heaton, R., & Hummel, J. (2019). Rapid unsupervised encoding of object files for visual reasoning. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 1895–1900). The Cognitive Science Society.

Henderson, J., & Ferreira, F. (1993, June). Eye movement control during reading: fixation measures reflect foveal but not parafoveal processing difficulty. *Canadian journal of experimental psychology = Revue canadienne de psychologie experimentale*, *47*(2), 201—221. doi: 10.1037/h0078814

Henderson, J. M. (2003). Human eye movements in visual search: A review. *Visual Cognition*, *10*(6), 471–494.

Hollenstein, R., & K., Z. (2019). Eye-tracking for named entity recognition. In *Proceedings of NAACL.*

Holsanova, J., Rahm, H., & Holmqvist, K. (2006). Entry points and reading paths on the newspaper spread: Comparing semiotic analysis with eye-tracking measurements. *Visual Communication*, *5*(1), 65–93.

Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., & Jawahar, C. V. (2021). ICDAR2019 competition on scanned receipt OCR and information extraction. *CoRR*, *abs/2103.10213*.

Jaume, G., Ekenel, H. K., & Thiran, J. (2019). FUNSD: A dataset for form understanding in noisy scanned documents. In *2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019, Sydney, Australia, September 22-25, 2019* (pp. 1–6). IEEE. doi: https://doi.org/10.1109/ICDARW.2019.10029

Lukander, K. (2016). User experience and interaction in eye-tracking studies. *International Journal of Human-Computer Interaction*, *32*(5), 343-350. doi: 10.1080/10447318.2016.1192064

Mathew, M., Bagal, V., Tito, R. P., Karatzas, D., Valveny, E., & Jawahar, C. V. (2021). InfographicVQA. *CoRR*, *abs/2104.12756*.

Mathews, A. P., Xie, L., & He, X. (2016). SentiCap: Generating image descriptions with sentiments. In D. Schuurmans & M. P. Wellman (Eds.), *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA* (pp. 3574–3580). AAAI Press. doi: 10.1609/AAAI.V30I1.10475

Migliaccio, A., MacDougall, H., Minor, L., & Della Santina, C. (2005). Eye movements in neuro-otological disorders: A review. *Journal of Neuro-Ophthalmology*, *25*(3), 232-239. doi: 10.1097/01.wno.0000182881.49778.a7

Nakagawa, M. (2015). Efficient sentence representation for statistical machine translation. In *Proceedings of the 53rd*

Annual Meeting of the Association for Computational Linguistics (ACL-2015) (pp. 1136–1145). Beijing, China.

Neubig, G., Dyer, C., Yoshino, K., & Matsumoto, Y. (2012). Inducing an efficient syntactic parser with large-scale data and probabilistic models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-2012)* (pp. 103–111). Jeju Island, Korea.

Noland, R., Weiner, M., Gao, Z., Cook, T., & Nelessen, A. (2017). Transportation and human factors: Using eye-tracking to improve safety. *Transportation Research Part F: Traffic Psychology and Behaviour*, *45*, 156-167. doi: 10.1016/j.trf.2016.09.004

Obaidellah, M., Al Haek, T., & Cheng, L. (2018). Eye-tracking in sports: Performance analysis and cognitive load. *Journal of Sports Sciences*, *36*, 28-34. doi: 10.1080/02640414.2017.1307179

OpenAI. (2023). *ChatGPT (mar 14 version).* `https://openai.com/chatgpt`. (Large Language Model)

Park, H., Subramaniyam, A., Hong, S., Kim, J., & Yu, H. (2017). Psycholinguistics of eye-tracking in healthcare: Applications and insights. *Journal of Health Communication*, *22*, 134-141. doi: 10.1080/10810730.2017.1284567

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244.

Penzkofer, A., Shi, L., & Bulling, A. (2024). Vsa4vqa: Scaling a vector symbolic architecture to visual question answering on natural images. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of NeurIPS 2021*.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124 3*, 372-422.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The e-z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476. doi: 10.1017/S0140525X03000104

Salvucci, D. D. (2001). The emma model: Linking cognitive attention to eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1278–1296.

Schilling, H., Rayner, K., Chace, K., Swaab, T., & Liversedge, S. P. (2006). Eye movements and lexical ambiguity resolution in reading. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1385-1399. doi: 10.1037/0096-1523.32.6.1385

Sharafi, Z. (2015). Product design and consumer behavior: The role of eye-tracking technology. *Design Studies*, *39*, 110-125. doi: 10.1016/j.destud.2015.01.003

Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 2, p. 629-633). doi: 10.1109/ICDAR.2007.4376991

Sood, E., Shi, L., Bortoletto, M., Wang, Y., Müller, P., & Bulling, A. (2023). Improving neural saliency prediction with a cognitive model of human visual attention. In *Annual Meeting of the Cognitive Science Society.*

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models.*

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97-136. doi: 10.1016/0010-0285(80)90005-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*.

Wang, H., Wang, Q., Li, Y., Wang, C., Chu, C., & Wang, R. (2023). DocTrack: A visually-rich document dataset really aligned with human eye movement for machine reading. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5176–5189). Association for Computational Linguistics.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., ... Lin, J. (2024). Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z., Xu, Y., Cui, L., Shang, J., & Wei, F. (2021). Layoutreader: Pre-training of text and layout for reading order detection. In *Conference on Empirical Methods in Natural Language Processing.*

Yan, X., Zhang, Y., & Zhang, C. (2024). Utilizing cognitive signals generated during human reading to enhance keyphrase extraction from microblogs. *Information Processing & Management*, *61*(2), 103614. doi: https://doi.org/10.1016/j.ipm.2023.103614

Zhang, J., Wang, H., & Luo, X. (2022). Dual-vie: Dual-level graph attention network for visual information extraction. In *PRICAI 2022: Trends in Artificial Intelligence* (pp. 422–434).

Zhang, Y., Li, Q., Nahata, S., Jamal, T., Cheng, S.-K., Cauwenberghs, G., & Jung, T.-P. (2024). Integrating Large Language Model, EEG, and Eye-Tracking for Word-Level Neural State Classification in Reading Comprehension. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *32*, 3465-3475. doi: 10.1109/TNSRE.2024.3435460