Human Eye-Tracking for Driving Explainability in DocVQA

Document Visual Question Answering

Marta Llopart Enajas

June 30, 2025

Abstract

This project addresses the lack of transparency in Document Visual Question Answering (DocVQA) models by comparing AI-generated attention with human cognitive processes captured via eye-tracking. It was recorded gaze data from 30 participants and performed causal occlusion experiments on a multimodal model to test its reliance on human-attended information. The experiments revealed that information prioritized by humans is both sufficient for the model to maintain its 70.7% accuracy and necessary for its high performance. These findings expose a critical "explainability gap," validating the use of human cognitive data as a ground truth for developing more transparent and trustworthy AI systems.

Keywords: Document Visual Question Answering (DocVQA), Explainable AI (XAI), Human Eye-Tracking, Attention

1 INTRODUCCTION

1.1 Context and Motivation

Document Visual Question Answering (DocVQA) is a crucial and complex task at the intersection of computer vision and natural language processing, designed to answer questions based on the contend of document images. The applications for this technology are wide-ranging and impactful, including automated contract analysis, invoice processing, and the digitization of historical documents. While current DocVQA models often achieve high levels of accuracy, their utility is frequently compromised by a lack of transparency in their decision-making processes. These AI models, which rely on sophisticated deep learning techniques, struggle to justify their outputs in a manner that aligns with human reasoning, creating a significant barrier to user trust and adoption.

To address this, various explainability techniques such as Grad-CAM, LIME, and Attention Maps have been developed to offer a window into the model's inner workings. However, a fundamental question remains unanswered: how well do these machine-generated explanations reflect actual human cognitive processes? This project addresses this challenge directly, aiming to bridge the gap between AI-generated attention and human-centered reasoning.

1.2 Project Objectives and Contributions

The primary goal of this project is to improve the interpretability and trustworthiness of DocVQA models by incorporating insights from human cognitive strategies. To achieve this, we utilized a Tobii Pro Spark eye-tracker to record the visual behavior of human participants as they engaged in DocVQA tasks, capturing their fixation patterns, search sequences, and even keystroke timings.

The main objectives of this thesis were to:

- Analyze and understand human attention by identifying which sections of a document participants focus on when answering questions.
- Directly compare human gaze data with AI-generated explainability maps to systematically identify alignments, discrepancies, and biases between the two.
- Explore how human perceptual strategies can be leveraged to improve the explanations generated by AI models, making them more intuitive and transparent.

By aligning AI attention mechanisms with observable human cognitive patterns, this project makes a significant contribution toward developing DocVQA systems that are not only accurate but also fundamentally more transparent, explainable, and trustworthy.

• Contact E-mail: marta.llopart@autonoma.cat
• Supervised by: Andrey Barsky (Ciències de la Computació)
• Academic Year 2024/25

1.3 Document Structure

The remainder of this document is organized as follows: Section 2 provides an overview of the state-of-the-art in DocVQA models and AI explainability techniques. Section 3 details the methodology used for the experimental setup, data collection, and analysis. Section 4 presents the final results of the project, including a comparative analysis of human and AI attention. Finally, Section 5 summarizes the main conclusions and proposes potential directions for future work.

2 State-of-the-Art

This project explores the collaboration between Document Visual Question Answering (DocVQA), AI explainability, and human cognitive science. This section reviews the relevant literature in these domains, establishing the technical background and motivation for incorporating human eye-tracking to enhance AI interpretability.

2.1 Document Visual Question Answering Models

DocVQA systems are designed to understand and answer questions about document images by processing both textual and visual information. These models typically rely on sophisticated deep learning architectures, such as Transformers, which have become the standard for multimodal document understanding.

A widely adopted approach in this domain involves adapting powerful language models, such as the Text-to-Text Transfer Transformer (T5), for multimodal tasks. These advanced models are capable of fusing three distinct types of information: they use a language backbone for semantic embeddings (the meaning of words), incorporate spatial embeddings derived from the bounding box coordinates of text (the layout of the page), and process the document scan with visual embeddings (the underlying image features). By combining these semantic, spatial, and visual inputs, the models can perform complex reasoning that accounts for both the content and the structure of the document. Although these architectures have achieved notable improvements in accuracy, their internal decision-making processes remain mostly inaccessible, a limitation that motivates this project.

2.2 Explainability in AI (XAI)

Explainability techniques aim to make the "black box" nature of deep learning models more transparent. In the context of VQA, methods like Grad-CAM, LIME, and attention maps are commonly used to generate visualizations that highlight the regions of an image a model "looks at" when forming an answer. However, these approaches are fundamentally model-centric and were not designed with human cognitive strategies in mind. This creates a critical gap: while these maps offer a form of justification, there is no guarantee that they reflect a reasoning process that is intuitive or understandable to a human user. Research has begun to question the alignment of these models with human perception, asking directly, "Do Multimodal Large Language Models See Like Humans?"

2.3 Eye-Tracking for Cognitive and AI Model Analysis

Eye-tracking technology provides a direct and objective method for observing human attention and cognitive processes during a task. It offers a unique opportunity to validate and improve AI systems by connecting them to observable human behavior. Although eye-tracking and explainable AI (XAI) have been explored independently, few studies have combined them in the context of document understanding.

Furthermore, researchers have examined whether humans and AI models focus on the same regions during Visual Question Answering (VQA) tasks, finding a significant discrepancy between human attention and model-generated attention maps [1]. More recent work has focused on bridging this gap by incorporating human-like attention to enhance VQA explainability [2, 3]. Similarly, the Voila-A framework aligns vision-language models with user gaze, which has been shown to improve both performance and interpretability [4].

Moreover, the development of new datasets such as VISTA highlights the increasing value of comprehensive visual and textual attention data for interpreting multimodal models [5]. Subtle aspects of human behavior, including strategies for managing distractor information [6], offer valuable insights for creating more robust and human-like AI systems. By collecting and analyzing human eye-tracking data specifically within the DocVQA domain, this project advances the field by aiming to produce AI systems whose explanations are not only generated, but also cognitively aligned with human users.

3 Methodology

The project followed a structured methodology consisting of four main phases, starting with the initial experimental design and setup through to data collection, analysis, and finally, comparison with the AI model. The approach ensured a robust and replicable workflow for collecting high-quality human attention data for DocVQA tasks.

3.1 Phase 1: Experimental Framework Development

The starting point of this project was the development of a robust experimental framework. The initial phase involved adapting an existing Python-based eye-tracking system to meet the specific requirements of DocVQA tasks. To optimize the experimental process, a new graphical user interface (GUI) was created using Tkinter. This custom GUI managed all stages of a participant's session and provided:

- Multilingual consent forms in Catalan, Spanish, and English
- Seamless integration with the Tobii Pro Spark eye-tracker for precise calibration and gaze tracking

- A user-friendly interface for conducting document-based question-answering trials
- Reliable data management, storing all recorded data - including gaze coordinates, timestamps, keystrokes, and responses — in a structured JSON format for each participant

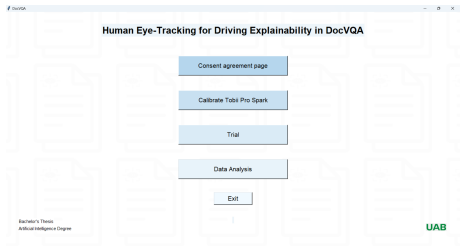

Fig. 1: GUI

Prior to full-scale data collection, three pilot sessions were conducted with volunteers. These sessions were essential for testing calibration accuracy, system robustness, and overall usability. Feedback from the pilots led to important refinements of the GUI, ensuring a more reliable and user-friendly environment for the final data collection phase.

3.2 Phase 2: Human Data Collection and Experimental Procedure

This phase was the core of the project, focusing on the careful execution of the eye-tracking trials with human participants.

3.2.1 Participant Recruitment

A total of 30 participants were successfully recruited for the study, exceeding the initial target and providing greater statistical power. Recruitment was conducted through a broad, informal network, including university classmates, friends, family, and community members. This strategy resulted in a diverse group of participants representing a wide range of ages and backgrounds, including students, teachers, and other working professionals.

The main requirement for participation was a sufficient understanding of English, as all documents and questions were presented in this language. To capture a wide range of natural viewing behaviors, no other exclusion criteria were applied; participants of all ages and with various types of vision (including those with and without glasses) were included. Participation was entirely voluntary. A small compensation was offered for the time committed to each trial, though several participants chose to contribute their time without accepting it.

3.2.2 Environment and Hardware Setup

The experimental setup was designed to be both controlled and flexible. Trials were conducted in various quiet rooms that provided good, consistent illumination to avoid interference with the eye-tracker.

The primary hardware consisted of a Huawei MateBook D14 (2020 edition) laptop, which ran the experimental software. The Tobii Pro Spark eye-tracker was magnetically mounted at the bottom of the laptop's screen, approximately 4-6 cm below the active display area. Participants were seated at a comfortable distance from the laptop. This distance was flexible to accommodate individual needs, particularly for participants who wore glasses. For those who needed to sit further back to achieve optimal focus, an external keyboard was provided so they could comfortably type their answers without reaching for the laptop's built-in keyboard.

3.2.3 Task Design and Experimental Walkthrough

The images used in the experiment were sourced from the official Single-Page Document Visual Question Answering (SP-DocVQA) dataset. From this large dataset, a subset of 40 document images was carefully selected for the trials. The selection was based on two criteria: visual clarity and orientation. Only landscape-oriented documents were chosen, as their larger format allowed the text to be displayed more clearly and legibly on the laptop screen.

The procedure for each session was carefully planned from start to finish:

1. **Onboarding and Consent:** At the start of each session, each participant was greeted and given a careful verbal explanation of the study's purpose and what the experiment would entail. They were then guided through the multilingual consent form presented in the GUI. The trial could only begin after they read the information and clicked a button to provide their consent. For participants who accepted compensation, the relevant personal data was collected separately.

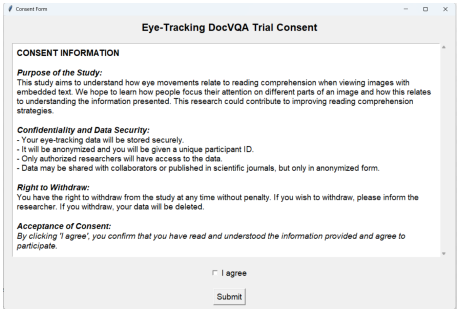

Fig. 2: Consent Agreement Page - GUI

2. **Calibration:** A precise calibration was crucial for data quality. Using the Tobii Pro Eye Tracker Manager software, a 5-point or 6-point calibration was performed. Participants were instructed to sit still and follow the points on the screen with their eyes. After the automated calibration, a manual validation step was performed: the participant was asked to look at several cross-shaped markers appearing on the screen. It was visually confirmed that the recorded gaze point was accurate. If any inaccuracies were detected, the calibration process was repeated until it was successful.
3. **Main Task:** The experiment began with an instruction screen explaining the task flow. For each trial, the

question appeared at the top of the screen, with the corresponding document image displayed directly below it. An answer box was situated at the bottom. Participants would read the question, search for the answer in the document, and type their response in the box. To proceed to the next trial, they pressed the "Enter" key. After each completed question, a "Good job!" message was displayed to provide positive reinforcement.

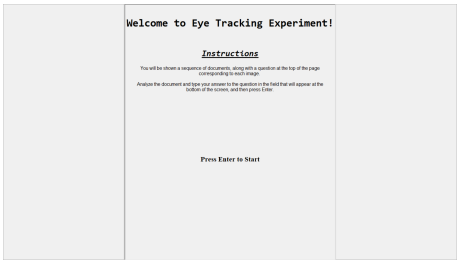

Fig. 3: Welcome Page

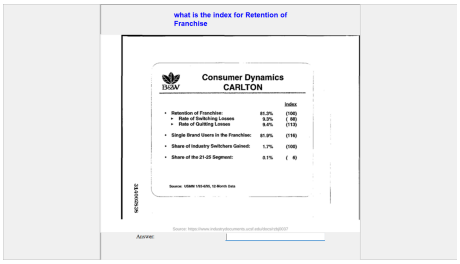

Fig. 4: Document Page

4. **Conclusion, Data Validation, and Debriefing:** After the final question was answered, a concluding screen appeared. Immediately following the trial, the *Data Analysis* was executed to ensure that the data had been collected correctly. Gaze heatmaps were generated for the just-completed session. As a final validation step, these heatmaps were shown to the participant, who could then informally corroborate that the visualizations accurately reflected their memory of where they had looked on the screen. The session concluded thanking the participant and holding a brief, informal chat to answer any remaining questions.

3.3 Phase 3: Data Preprocessing and Analysis

The raw data collected in the JSON files was then preprocessed for analysis. This step utilized a pre-existing script, which was adapted for the project's needs. The primary output of this phase was a set of visualizations where gaze heatmaps were overlaid onto the document images, providing an intuitive representation of where participants focused their attention.

3.3.1 Fixation Identification and Clustering

Raw eye-tracking data consists of a high-frequency sequence of gaze coordinates, which is too noisy for direct interpretation. The first crucial preprocessing step was to identify "fixations"—periods where a user's gaze remained stable, indicating a pause to process information.

To achieve this, the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm was employed. This algorithm grouped spatially-proximate gaze points (based on their X and Y coordinates) into distinct clusters. A `min_cluster_sample` hyperparameter was used to define the minimum number of data points required to form a valid fixation, effectively setting a minimum duration for a user's pause and filtering out rapid, unintended movements (saccades).

3.3.2 Behavioral Phase Separation

To gain deeper insight into the user's cognitive process, the analysis separated the gaze data into distinct behavioral phases. The timestamp of the user's first keystroke was used as the critical dividing line to distinguish between information gathering and answer formulation. This resulted in three separate datasets for each trial:

- **Pre-Typing Phase:** All gaze data occurring before the first keystroke, representing the user's reading, search, and comprehension process.
- **Typing Phase:** All gaze data occurring at or after the first keystroke, representing any visual reference while the user was typing their answer.
- **Last Pre-Typing Fixation:** The single, final fixation event that occurred immediately before the first keystroke. This was isolated to represent the user's last point of focus before committing to an answer.

3.3.3 Generation of Human Attention Maps

The ultimate goal was to create a quantitative "attention map" for each phase that could be directly compared to the output of the computational model. For each fixation identified in a given phase, the script located all OCR tokens on the document that fell within a predefined `fixation_radius`.

Each token received an attention score based on the number of fixations that landed on or near it. This *fixation count* method was chosen over a duration-based model to emphasize the locations the user repeatedly checked. The final data for each trial was saved to a consolidated JSON file, containing separate, normalized attention maps for each of the three behavioral phases.

3.3.4 Multi-Layered Visualizations

The initial heatmaps were improved to visualize this multi-phase data on a single plot. The final plots display a layered heatmap where each color represents a different phase, allowing for intuitive analysis of the user's entire decision-making process:

- **Blue Heatmap:** Represents the general "Pre-Typing" phase. The color is a gradient, becoming darker for fixations that occurred closer in time to when the user started typing.
- **Green Heatmap:** Represents the "Typing" phase, showing where the user looked while inputting their answer.

- **Red Outline:** Represents the crucial ”Last Pre-Typing Fixation,” highlighting the user’s final fixation with a distinctive border drawn on top of all other layers.

These plots also include a header displaying the question and a detailed footer showing the user’s answer, the correct answer, and a legend for the color scheme.

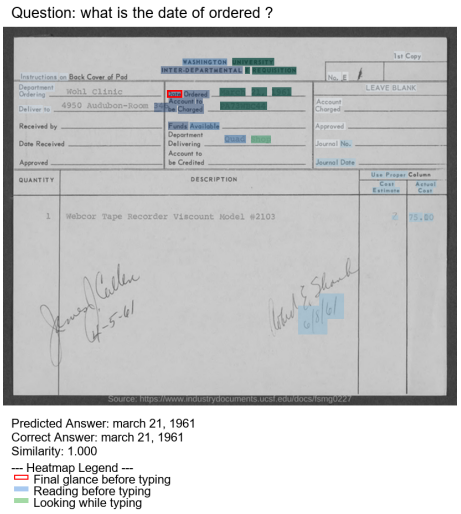

Fig. 5: Example Human Attention Heatmap

3.4 Phase 4: AI Model Comparison

The final phase of the methodology focused on the AI model. The model selected for this project was VT5, a multimodal transformer architecture that combines semantic, spatial, and visual embeddings to understand document content and layout.

The initial work involved testing the provided VT5 codebase on the standard validation set to confirm that the implementation worked as expected. It was observed that running the model efficiently required a GPU, which was not available during this phase of the project. As a result, subsequent tests were conducted using the much smaller dataset collected from the 30 participant sessions, which could be processed in a reasonable time without a dedicated GPU. The goal was to generate AI explainability maps from this model to compare against the human gaze data collected in Phase 2.

3.4.1 Initial Focus: Visual vs. Textual Attention

The initial explainability analysis focused exclusively on the model’s visual attention. The goal was to determine if the model’s focus on different spatial regions of the document image, represented by a heatmap over the visual patches, could explain its reasoning.

However, it was observed that these visual heatmaps were often too general and did not clearly correlate with the specific information required by the question. The model’s visual attention did not appear to be a sufficiently accurate or interpretable signal for this particular task.

Given the limitations of the strictly visual approach, the focus of the analysis shifted to the model’s textual attention. This method, which highlights specific OCR words and

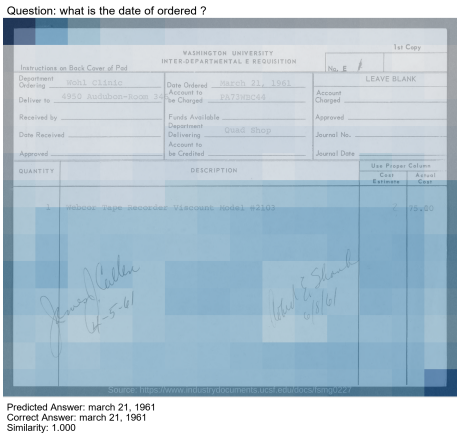

Fig. 6: Example Model Visual Attention Heatmap

phrases, provides a much more detailed and interpretable form of explainability. Importantly, this textual attention map is directly comparable to the human gaze data, which is also centered on reading text, thus enabling a direct and meaningful comparison.

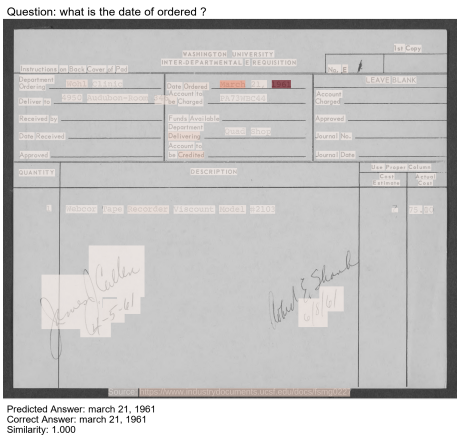

Fig. 7: Example Model Textual Attention Heatmap

3.4.2 Model Attention Map Generation

To create an explainability map that shows what the model is ”looking at,” the cross-attention scores from the final decoder layer of the VT5 transformer were extracted. These scores represent the level of focus the model places on each token of the input sequence as it generates its answer.

The model utilizes a multi-head attention mechanism, where each ”head” can learn to focus on different types of information. To simplify this complex, multi-layered output into a single, interpretable score for each input token, the maximum attention value across all heads was taken as the definitive score for that token.

3.4.3 Word-Level Attention Aggregation

A significant challenge in comparing model and human data is the mismatch between the model’s ”tokens” and human-readable ”words.” The model’s tokenizer often breaks single words from the OCR data into multiple smaller sub-word tokens (e.g., ”opportunity” might become `["opportun", "ity"]`).

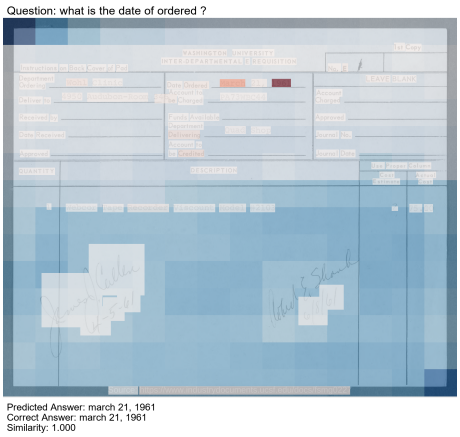

Fig. 8: Example Model Combined Attention Heatmap

To resolve this, a geometric aggregation method was implemented. For each original word and its corresponding bounding box from the OCR data, the script identifies all sub-word tokens whose calculated center point is physically located inside that word's box. The final attention score for the word is then determined by taking the maximum attention score from all the sub-tokens it contains. This process yields a clean, per-word attention map that is directly comparable to the human data. This word-level attention map, along with the model's predicted answer and the correct answer, was saved to a JSON file for each trial.

3.4.4 Quantitative Comparison Methodology

To perform a quantitative comparison between human and model attention, it was first necessary to establish robust human consensus heatmaps from the collective gaze data, which would serve as ground truth. This was accomplished by aggregating attention data from all participants for each unique document and categorizing the results into three baselines based on response correctness (`human_correct`, `human_semi_correct`, and `human_wrong`).

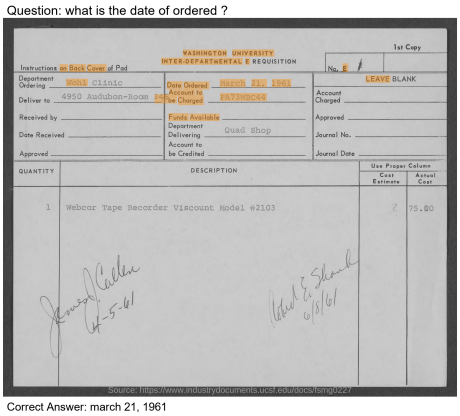

Fig. 9: Example Consensus Attention Heatmap

Within each category, the number of unique users who fixated on each specific OCR token was counted. A `CONSENSUS_THRESHOLD` was then applied: a bounding box or word was included in the final consensus map if the percentage of users who fixated on it met or exceeded this threshold. This approach filters out individual-specific gaze patterns, resulting in a stable representation of each group's collective attention.

With both the model's per-word attention maps and the multiple human consensus baselines established, a comprehensive quantitative comparison was performed using two distinct methods.

1. **Spearman's Rank Correlation:** To measure the global similarity in how the model and humans prioritize information across an entire document, the Spearman correlation coefficient was calculated. This metric was chosen for its robustness, as it evaluates the similarity in the *rank order* of attended words rather than their absolute attention values, making it unaffected by scaling differences between model and human attention scores. The resulting coefficients were then aggregated in two ways to reveal different patterns:
   - **By Model Performance:** To understand how attention alignment relates to the model's success, correlations were averaged based on the model's answer category (correct, semi_correct, wrong).
   - **By Question Type:** To identify if alignment varies by task, the correlations were also independently averaged by question category (ENTITY, NUMERIC, OTHER).
2. **Top-5 Attention Overlap:** To extend global correlation and provide a more direct measure of agreement on the most critical evidence, a *graduated Top-N overlap analysis* was conducted for N values of 3, 5, and 10. This method offers a detailed view of similarity, testing for agreement on the most important word (Top-3), the primary set of key evidence (Top-5), and the broader contextual area (Top-10).

   Furthermore, this analysis was performed as a *full matrix comparison*. Each of the three model performance categories was compared against each of the three human consensus baselines. This comprehensive approach enables a deep analysis of specific interactions, making it possible to determine, for example, whether a correctly performing model's attention is more aligned with a successful human's strategy or one associated with errors.

3.4.5 Causal Analysis via Occlusion Experiments

While the correlational metrics described previously can reveal an association between human and AI attention, they cannot establish a cause-and-effect link. To investigate whether the information prioritized by humans has a direct impact on the model's performance, a series of occlusion experiments were designed.

The methodology involved comparing the model's baseline performance on the original documents against its performance on precisely modified inputs where specific information was either revealed or removed. The impact of these manipulations was measured by the change in two key performance metrics: Accuracy and Average Normalized Levenshtein Similarity (ANLS).

Two distinct experiments were conducted to test for sufficiency and necessity.

Experiment 1: Testing for Sufficiency

This experiment was designed to answer the question: *Is the information attended to by humans sufficient for the model to perform its task correctly?*

To test this, a *human-only* version of each document was generated. This input was created on a completely black canvas of the same dimensions as the original document. The script then selectively revealed only the regions corresponding to the human consensus attention map. To maintain consistency between the visual and textual modalities, the input data for the model was also filtered to contain only the words and bounding boxes appearing in these visible regions.

A critical detail was the explicit preservation of the ground-truth answer's location. The bounding boxes for the answer were always revealed, even if they fell outside the human consensus map. This step was essential to ensure the task remained solvable.

This process created a new image where the model could only "see" what the human group collectively focused on, plus the answer itself. This modified image was then passed to the model for evaluation. Strong model performance under these conditions would suggest that the information prioritized by humans is indeed sufficient for task completion.

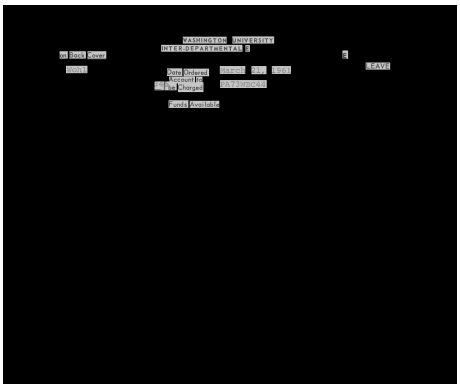

Fig. 10: Example Occlusion Image

Experiment 2: Testing for Necessity

Alternatively, this experiment aimed to answer the question: *Is the information attended to by humans necessary for the model's performance?*

To test this, the original document input was modified by occluding the regions identified by the human consensus map. Both the image regions and their corresponding text and bounding box information were removed from the model's input. As in the sufficiency experiment, maintaining the ground-truth answer was essential. Any part of the answer that overlapped with a human consensus region was explicitly excluded from occlusion, thereby isolating the effect of removing surrounding information.

This *context-denied* image was then passed to the model. A significant drop in the model's Accuracy and ANLS scores in this condition would provide strong evidence that the model relies on the same information as humans and that this information is therefore necessary for its reasoning process.

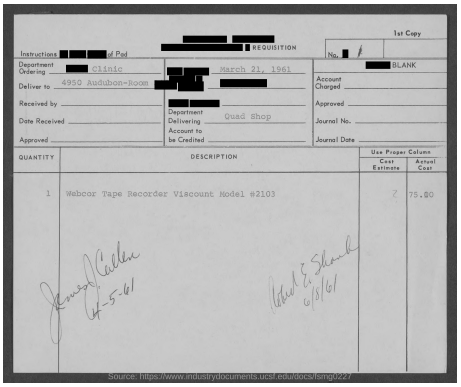

Fig. 11: Example Opposite Occlusion Image

4 Results

This section presents the detailed quantitative findings from the project. It is organized into two main parts. First, it analyzes the strategic alignment between human and model attention patterns across the three experimental conditions. Second, it evaluates the model's task performance (Accuracy and ANLS) under these same conditions to measure the causal impact of the experimental changes.

4.1 Analysis of Human-AI Attention Alignment

The first analysis focuses on the alignment between the model's attention and the human consensus baseline, with detailed results summarized in Table 1 and Table 2.

TABLE 1: Average Spearman's Rank Correlation (ρ) between Human and Model Attention. (n=number of samples)

Model Perf.	Baseline	Sufficiency (Occluded)	Necessity (Opposite)
Correct	0.1759 (n=55)	0.7144 (n=51)	-0.1871 (n=31)
Semi-Corr.	0.2053 (n=8)	0.6456 (n=17)	-0.2498 (n=14)
Wrong	0.0461 (n=8)	0.8303 (n=3)	-0.3929 (n=26)

TABLE 2: Top-N Attention Overlap (%) between Model (Correct) and Human (Correct) baselines. (n=number of samples)

Overlap Level	Baseline	Sufficiency (Occluded)	Necessity (Opposite)
Top-3	18.75% (n=32)	35.48% (n=31)	15.79% (n=19)
Top-5	19.38% (n=32)	40.00% (n=31)	14.74% (n=19)
Top-10	23.75% (n=32)	52.58% (n=31)	13.68% (n=19)

4.1.1 Baseline Human-AI Alignment (Standard Model)

First, the standard VT5 model was compared against the human consensus baseline without any occlusions to estab-

lish the default level of strategic alignment. The results of this baseline comparison are summarized in the 'Baseline' column of both tables.

- **Correlation Analysis:** As shown in Table 1, the baseline correlation is positive but weak when the model performs well (e.g., rho=0.1759 for *CORRECT* cases), indicating a slight positive relationship in how the model and humans rank the importance of words. This correlation decreases to near-zero for cases where the model was definitively WRONG.
- **Attention Overlap:** Table 2 reveals a similar pattern of low alignment. When comparing a *CORRECT* model against the Human Correct baseline, the overlap is consistently below 25%. This crucial result demonstrates that even when the model and humans both arrive at the correct answer, they achieve this by focusing on largely different sets of evidence, pointing to fundamentally distinct problem-solving strategies in the unrestricted, baseline condition.

4.1.2 Impact of Causal Interventions on Attention Alignment

Next, the two occlusion experiments were analyzed to determine if human-attended information is sufficient and/or necessary for the model, and how these interventions affect attention alignment.

Sufficiency Experiment (Occluded Model)

This experiment tested whether the information humans consider important is sufficient for the model by blacking out all other information, while explicitly preserving the answer region. The results, shown in the 'Sufficiency' column of the tables, show a significant shift in the model's behavior.

- **Correlation Analysis:** As detailed in Table 1, forcing the model to use only human-attended information caused the average correlation to increase dramatically from the baseline of 0.1759 to 0.7144 for the *CORRECT* model cases. This demonstrates that when constrained to the "human path," the model's internal ranking of information becomes highly aligned with human patterns.
- **Attention Overlap:** A similar increase is evident in Table 2. The Top-5 overlap for CORRECT model cases more than doubled, from 19.38% to 40.00%, with the Top-10 overlap exceeding 50%. This confirms that the model can effectively replicate a human-like attention strategy when its information access is limited to what humans use.

Necessity Experiment (Opposite Occluded Model)

This experiment tested whether the model depends on human-attended information by blacking out the Human Correct consensus regions, while explicitly preserving the answer area. The results are presented in the final column of each table.

- **Correlation Analysis:** Removing the information that humans use caused the attention correlation to become negative across all categories (e.g., -0.1871 for the *CORRECT* model cases). This indicates that the model, when forced to find alternative evidence, focuses on information that humans actively ignored, creating a significant and measurable strategic misalignment.
- **Attention Overlap:** The Top-N overlap analysis further quantifies the strategic misalignment caused by occluding human-preferred evidence. As shown in Table 2, the overlap between a CORRECT model and the Human Correct baseline dropped across all levels, with the Top-5 overlap falling to 14.74% from a baseline of 19.38%. This decrease confirms that removing the primary evidence used by humans forces the model into a less human-like strategy.

  The small, residual overlap that remains is explained by the experimental design's "answer preservation" rule, which ensures the ground-truth answer is never occluded to keep the task solvable. Therefore, this small overlap represents the shared, minimal focus on the answer text itself, while the significant drop from the baseline demonstrates that the model's alignment on relevant contextual cues is broken when the primary human-attended evidence is removed.

4.2 Model Performance Analysis

The second part of the analysis evaluated the impact of the occlusion experiments on the model's performance. The goal was to determine whether providing or removing human-attended information would cause a measurable change in the model's Accuracy and Average Normalized Levenshtein Similarity (ANLS). The results can be seen in Table 3.

TABLE 3: Model Performance (Accuracy and ANLS) Across All Experimental Conditions.

Evaluation Condition	Mean Accuracy	Mean ANLS
Full Dataset Baseline	55.0%	0.661
Human Subset (All Qs)	75.1%	0.851
Human Subset (1st Q)	**70.7%**	**0.841**
Sufficiency (Human-view)	70.7%	0.878
Necessity (Opp. Human-view)	46.3%	0.605

4.2.1 Baseline Performance

First, to establish a clear point of reference, the model's performance was measured on three different subsets of the data without any occlusion:

- On the full, large validation dataset, the model achieved a Mean Accuracy of 55.0% and a Mean ANLS of 0.661.
- When evaluated on the specific subset of documents used in the human experiments (evaluating all questions), the baseline performance was significantly higher, with a Mean Accuracy of 75.1% and a Mean ANLS of 0.851. This significant performance improvement suggests that the subset of documents selected for the human experiments were, on average,

less complex or contained less ambiguity than the full validation set, making them a high-quality sample for analysis.

- To ensure a fair comparison for the subsequent experiments, a final baseline was established by evaluating only the first question for each unique document. This more rigorous "first-look" baseline resulted in a Mean Accuracy of 70.7% and a Mean ANLS of 0.841. This is the primary baseline used for comparison.

4.2.2 Performance on Occluded Inputs

The two causal experiments produced distinct and informative changes in the model's performance when compared to the *first-look* baseline.

Experiment 1: Sufficiency (Human-View Only)

Under the *sufficiency* condition, where the model's input was restricted to only human-attended regions (plus the answer), the model's performance was remarkably stable as seen in Table 3:

- *Mean Accuracy:* 70.7%
- *Mean ANLS:* 0.878

This result is highly significant. The model's accuracy remained identical to the baseline, while the ANLS score saw a notable increase of 0.037. This finding demonstrates that the information humans attend to is not only sufficient for the model to perform its task but may even help it generate more precise and structurally correct answers, as indicated by the improved ANLS. The model does not need access to the rest of the document to maintain its top-line accuracy.

Experiment 2: Necessity (Hiding Human-View)

Conversely, the *necessity* experiment, which tested if the model required human-attended context by occluding those regions, caused a dramatic drop in performance:

- *Mean Accuracy:* 46.3%
- *Mean ANLS:* 0.605

The model's accuracy dropped by 24.4 points from the baseline, and the ANLS score decreased significantly by 0.236. This provides strong evidence that the contextual information attended to by humans is indeed necessary for the model's reasoning process. Without these primary indicators, the model's ability to locate and formulate the correct answer was substantially reduced, confirming that the model and humans rely on a similar set of contextual evidence to solve the task.

5 Conclusions

This project aimed to investigate the gap between human cognition and AI reasoning in the complex task of Document Visual Question Answering (DocVQA). By using eye-tracking technology to capture human attention patterns, this thesis has made several key contributions to the field of explainable AI. The analysis produced a clear and significant finding: a fundamental divergence exists between how humans and the AI model approach the same task. The initial baseline comparison revealed that even when both the model and a human arrive at the same correct answer, their underlying methods are dissimilar. This is demonstrated by a weak correlation in their attention patterns, with a Spearman's rank correlation of just 0.1759 for correct cases, and a low Top-5 attention overlap of 19.38%.

To investigate this relationship further, a series of causal occlusion experiments were conducted to examine the connection between human and model attention in greater depth. The findings from these experiments provide two critical insights. First, when the model was restricted to only the regions attended by humans, its accuracy remained stable at 70.7%, and its answer precision (ANLS) even improved. This demonstrates that the information prioritized by humans is not only sufficient for the model to solve the task, but may also enhance its precision. Second, when these same human-attended regions were removed, the model's performance dropped significantly, with accuracy falling to 46.3%. This provides strong evidence that access to human-prioritized information is also necessary for the model's reasoning process.

Together, these results demonstrate that both the model and humans depend on a similar set of contextual elements to achieve high performance, even though their natural attention strategies may differ. However, these findings also reveal an important challenge: the "explainability gap." While the model can reach human-level performance when guided to focus on human-attended regions, its internal attention mechanisms do not naturally align with human cognitive patterns. This misalignment means that explanations generated solely from model attention can be misleading, failing to meet the transparency and trustworthiness required for truly explainable AI systems. These findings highlight the importance of using human cognitive data as a ground truth for developing AI models that are not just accurate, but also fundamentally aligned with how users reason and interpret information.

These findings, in turn, open up several promising directions for future research. While this project successfully conducted its analysis on the dataset collected from human participants, a crucial next step would be to scale these causal experiments to larger reference datasets to determine whether the findings hold true in general. Furthermore, the rich dataset of human attention should be utilized not just for comparative analysis, but also for active model improvement. Future work could explore methods for integrating the collected human gaze data directly into the training or fine-tuning process of a DocVQA model, building on existing research that aims to align AI with human attention. A key assumption is that human-aligned explanations are more trustworthy, which should be formally tested through user studies that quantitatively measure user trust and understanding. Finally, this analysis could be extended to other model architectures to investigate whether the explainability gap is a general phenomenon across the field of multimodal AI.

6 GitHub

All the code and results can be seen on GitHub

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to the people and institutions who made this project possible.

To begin with, I wish to express my deepest appreciation to my tutor, Andrey Barsky, for his invaluable guidance, support, and supervision throughout this entire project. My sincere thanks go to the Computer Vision Center (CVC) for providing the essential material for my research, specifically the Tobii Pro Spark eye-tracker that was central to my data collection.

I am deeply grateful to all 30 participants—my friends, family, classmates, and community members—who generously volunteered their time to help me obtain the data necessary for this thesis. I would also like to thank my coordinator for putting me in contact with the person who provided me with GPU access. Additionally, I offer my thanks to the administrative worker from the CVC who managed the necessary documents to ensure participants could be compensated for their time.

Finally, a special thank you to my friends and family for their constant encouragement and support during these months of hard work. This project would not have been completed without all of their contributions.

REFERENCES

[1] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017, pp. 932–937.

[2] G. Liu, J. Zhanga, A. B. Chan, and J. H. Hsiao, "Enhancing document vqa explainability through human-guided attention," *arXiv preprint arXiv:2305.03601*, 2023.

[3] E. Sood, F. Kögel, P. Müller, D. Thomas, M. Bace, and A. Bulling, "Multimodal integration of human-like attention in visual question answering," *arXiv preprint arXiv:2109.13139*, 2023.

[4] K. Yan, L. Ji, Z. Wang, Y. Wang, N. Duan, and S. Ma, "Voila-a: Aligning vision-language models with user's gaze attention," *arXiv preprint arXiv:2401.09454*, 2024.

[5] Harshit and T. Tasdize, "Vista: A visual and textual attention dataset for interpreting multimodal models," *arXiv preprint arXiv:2410.04609*, 2024.

[6] X. Xu and H. Chen, "Human-like distractor response in vision-language model," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, Eds. Nusa Dua, Bali: Association for Computational Linguistics, Nov. 2023, pp. 174–185. [Online]. Available: https://aclanthology.org/2023.ijcnlp-main.12/

[7] E. Kasneci, H. Gao, S. Ozdel, V. Maquiling, E. Thaqi, C. Lau, Y. Rong, G. Kasneci, and E. Bozkir, "Introduction to eye tracking: A hands-on tutorial for students and practitioners," *arXiv preprint arXiv:2404.15435*, 2024.

[8] J. Lin, S. Ye, and R. W. Lau, "Do multimodal large language models see like humans?" *arXiv preprint arXiv:2412.09603*, 2025.

APPENDIX

A.1 Results Human-AI Attention

TABLE 4: Full Overlap Matrix for the Baseline (No Occlusion) Experiment (%). Each cell shows the average attention overlap between a model performance category (rows) and a human consensus baseline (columns). (n=number of samples)

		Human Consensus Baseline Category		
Overlap Level	**Model Performance**	**Human Correct**	**Human Semi-Correct**	**Human Wrong**
Top-3	Model Correct	18.75% (n=32)	4.17% (n=16)	14.29% (n=7)
	Model Semi-Correct	16.67% (n=4)	0.00% (n=3)	33.33% (n=1)
	Model Wrong	8.33% (n=4)	0.00% (n=2)	0.00% (n=2)
Top-5	Model Correct	19.38% (n=32)	13.75% (n=16)	8.57% (n=7)
	Model Semi-Correct	10.00% (n=4)	6.67% (n=3)	20.00% (n=1)
	Model Wrong	5.00% (n=4)	0.00% (n=2)	10.00% (n=2)
Top-10	Model Correct	23.75% (n=32)	26.25% (n=16)	18.57% (n=7)
	Model Semi-Correct	22.50% (n=4)	23.33% (n=3)	20.00% (n=1)
	Model Wrong	2.50% (n=4)	5.00% (n=2)	10.00% (n=2)

TABLE 5: Full Overlap Matrix for the Sufficiency (Occluded) Experiment (%) (n=number of samples).

		Human Consensus Baseline Category		
Overlap Level	**Model Performance**	**Human Correct**	**Human Semi-Correct**	**Human Wrong**
Top-3	Model Correct	35.48% (n=31)	7.14% (n=14)	11.11% (n=6)
	Model Semi-Correct	42.86% (n=7)	0.00% (n=6)	16.67% (n=4)
	Model Wrong	16.67% (n=2)	0.00% (n=1)	N/A (n=0)
Top-5	Model Correct	40.00% (n=31)	21.43% (n=14)	16.67% (n=6)
	Model Semi-Correct	51.43% (n=7)	10.00% (n=6)	25.00% (n=4)
	Model Wrong	40.00% (n=2)	40.00% (n=1)	N/A (n=0)
Top-10	Model Correct	52.58% (n=31)	33.57% (n=14)	25.00% (n=6)
	Model Semi-Correct	48.57% (n=7)	36.67% (n=6)	25.00% (n=4)
	Model Wrong	70.00% (n=2)	70.00% (n=1)	N/A (n=0)

TABLE 6: Full Overlap Matrix for the Necessity (Opposite Occluded) Experiment (%).

		Human Consensus Baseline Category		
Overlap Level	**Model Performance**	**Human Correct**	**Human Semi-Correct**	**Human Wrong**
Top-3	Model Correct	15.79% (n=19)	4.17% (n=8)	16.67% (n=4)
	Model Semi-Correct	4.76% (n=7)	5.56% (n=6)	0.00% (n=1)
	Model Wrong	0.00% (n=14)	4.76% (n=7)	6.67% (n=5)
Top-5	Model Correct	14.74% (n=19)	7.50% (n=8)	15.00% (n=4)
	Model Semi-Correct	11.43% (n=7)	10.00% (n=6)	20.00% (n=1)
	Model Wrong	0.00% (n=14)	8.57% (n=7)	4.00% (n=5)
Top-10	Model Correct	13.68% (n=19)	10.00% (n=8)	15.00% (n=4)
	Model Semi-Correct	15.71% (n=7)	30.00% (n=6)	10.00% (n=1)
	Model Wrong	3.57% (n=14)	14.29% (n=7)	6.00% (n=5)