A Sentence-Level Visualization of Attention in Large Language Models

Seongbum Seo
Sejong University
seo@seongbum.com

Sangbong Yoo
Sejong University
usangbong@gmail.com

Hyelim Lee
Korea University
hyelim_lee@korea.ac.kr

Yun Jang*
Sejong University
jangy@sejong.edu

Ji Hwan Park
Rochester Institute of Technology
jpark@mail.rit.edu

Jeong-Nam Kim
The University of Oklahoma
layinformatics@gmail.com

Abstract

We introduce **SAVIS**, a sentence-level attention visualization tool that enhances the interpretability of long documents processed by Large Language Models (LLMs). By computing inter-sentence attention (ISA) through token-level attention aggregation, **SAVIS** reduces the complexity of attention analysis, enabling users to identify meaningful document-level patterns. The tool offers an interactive interface for exploring how sentences relate to each other in model processing. Our comparative analysis with existing visualization tools demonstrates that **SAVIS** improves task accuracy and reduces error identification time. We demonstrate its effectiveness for text analysis applications through case studies on various analysis tasks. **SAVIS** is available at https://pypi.org/project/savis with a screencast video at https://youtu.be/fTZZPHA55So.

1 Introduction

Attention mechanisms in language models enable interpretation of how models process text by showing weights assigned to different input elements (Zhao et al., 2024). Recent Large Language Models (LLMs) like GPT-4 (OpenAI et al., 2024) and PaLM 2 (Anil et al., 2023) have achieved state-of-the-art performance across various natural language processing (NLP) tasks, using these attention mechanisms as core components. However, interpreting attention patterns becomes increasingly challenging as documents grow longer, particularly when analyzing how models process document-level context.

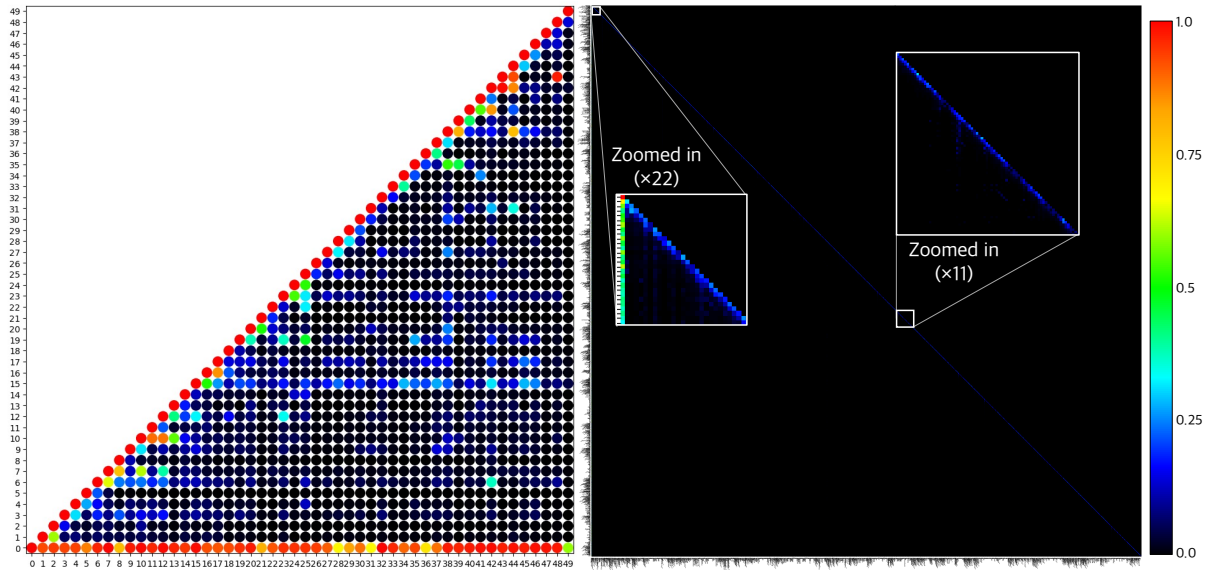
The challenge of interpretation scales with both model size and input length. Transformer-based models like BERT contain 144 attention heads across 12 layers (Clark et al., 2019), each learning distinct patterns ranging from syntactic dependencies to broader semantic relationships. For practitioners without deep NLP expertise, these patterns are complicated to interpret as they must examine hundreds of attention patterns simultaneously. Studies show that only certain heads serve important functions, with one study finding that only 10 out of 48 encoder heads are sufficient to maintain translation quality in machine translation models (Voita et al., 2019). Michel et al. (Michel et al., 2019) found that up to 60% of attention heads can be pruned without significantly impacting model performance.

Various approaches have been proposed to visualize attention in LLMs, from early attention-matrix heatmaps (Bahdanau et al., 2015; Rush et al., 2015) to bipartite graph representations (Lee et al., 2017; Liu et al., 2018). More recent tools include BertViz (Vig, 2019) for multi-scale visualization, LIT (Tenney et al., 2020) and Dodrio (Wang et al., 2021) for interactive analysis, and KnowledgeVIS (Coscia and Endert, 2023) for semantic exploration. However, these token-level visualization methods face a fundamental scalability challenge: for a document with n tokens across l layers and h attention heads per layer, practitioners must examine $l \times h \times n^2$ attention connections.

To address this challenge, we introduce **SAVIS**, an attention visualization tool that enhances the interpretability of long documents processed by LLMs. Through case studies and comparative analysis, we demonstrate that our sentence-level approach significantly reduces the time required to identify attention patterns while maintaining interpretability. Our contributions are (1) a method for aggregating token-level attention at the sentence level for a more straightforward interpretation of document-level patterns and (2) an interactive visualization tool for exploring sentence relationships, with features designed to reveal how models process sequential information.

*Corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 313–320



(a) Inter-Sentence Attention Visualization (b) Token-Level Attention Heatmap

Figure 1: Visualization of attention patterns in the Wikipedia article *'Computational linguistics'* (https://en.wikipedia.org/wiki/Computational_linguistics), using the same data and color scale for both visualizations. (a) Our proposed inter-sentence attention visualization for 50 sentences shows clear attention patterns between key sentences. (b) Traditional token-level attention heatmap for 2,233 tokens, demonstrating the difficulty in identifying specific areas of interest in large documents.

2 Inter-Sentence Attention

Processing long documents with transformer-based models requires understanding the relationships between sentences. However, interpreting these relationships through token-level attention patterns is challenging due to their complexity. We introduce inter-sentence attention (ISA), which captures sentence-level relationships by aggregating token-level attention patterns.

Given two sentences S_a and S_b with token indices $[i_a, i_{a+1})$ and $[i_b, i_{b+1})$ respectively, we compute their inter-sentence attention through a three-step process. First, we integrate attention patterns across layers by taking the maximum attention score at each position: $A(i,j) = \max_{l \in L} \alpha_l(i,j)$, where $\alpha_l(i,j) = \text{softmax}(Q_{l,i}K_{l,j}^T/\sqrt{d_k})V_{l,j}$ is the standard scaled dot-product attention at layer l. Here, $Q_{l,i}$ and $K_{l,j}$ are query and key vectors for tokens i and j in layer l, $V_{l,j}$ is the value vector, and d_k is the dimension of key vectors. We then compute attention scores between sentence pairs by taking the maximum attention score between any token pair in two sentences: $\beta_h(S_a, S_b) = \max_{(i,j) \in S_a \times S_b} A(i,j)$ for each attention head h. Finally, we aggregate these scores across attention heads by selecting the maximum value: $\text{ISA}(S_a, S_b) = \max_{h \in H} \beta_h(S_a, S_b)$, where H is the set of attention heads.

While prior work has used averaging across attention heads (Abnar and Zuidema, 2020), we opt for taking the maximum attention score to preserve strong signals from individual heads. This choice is motivated by findings that different attention heads often specialize in capturing specific linguistic patterns (Clark et al., 2019). This aggregation approach first integrates attention patterns across layers and then computes sentence-level attention by identifying relationships between sentence pairs. The process reduces the computational complexity from $O(n^2)$ to $O(m^2)$ for a document with n tokens and m sentences, where $n \gg m$ since each sentence typically consists of multiple tokens. This reduction enables analysis of document-level attention patterns while maintaining the most salient connections between sentences.

3 Visualization Tool

We present **SAVIS**, an open-source Python library built with `matplotlib` and `transformers` for visualizing attention patterns at multiple scales in LLMs. The library is available at https://pypi.org/project/savis. Implemented in Python for Jupyter Notebook environments, **SAVIS** supports transitions between sentence-level and token-level visualizations.

3.1 Inter-Sentence Attention Visualization

The primary interface presents attention patterns through a dot plot visualization (Figure 1 (a)). Each point represents an attention relationship between two sentences, with color intensity indicating attention strength. Users can examine specific relationships through a hovering mechanism by viewing the relevant sentences and their attention scores. The following code demonstrates the sentence-level visualization:

```
from savis import TextGenerator, ISA,
    ISAVisualization

# Generate text and extract attention
generator = TextGenerator("<model_name>")
text, attentions, tokenizer, _, outputs = \
   generator.generate_text(input_text)

# Compute inter-sentence attention
isa = ISA(outputs.sequences[0], attentions,
    tokenizer)

# Create interactive visualization
vis = ISAVisualization(
   sentence_attention=isa.sentence_attention,
   sentences=isa.sentences)
vis.visualize_sentence_attention()
```

3.2 Multi-Scale Analysis

SAVIS provides three complementary views for analyzing attention at different scales. The sentence-level view offers a high-level overview of document structure through ISA patterns. This approach is efficient for long documents where token-level visualization becomes overwhelming. As shown in Figure 1, when analyzing a Wikipedia article with 2,233 tokens, our inter-sentence visualization reveals clear patterns between key sections, while the token-level heatmap becomes difficult to interpret due to the dense number of connections. To examine specific attention patterns in detail, users can select any pair of sentences and visualize token-level attention weights between them. This focused token-level view helps understand how attention flows between specific parts of the document:

```
# Visualize full document token attention
vis.visualize_token_attention_heatmap(
   attentions, tokenizer, input_ids,
   layer=-1  # Specify layer
)

# Focus on specific sentence pair
vis.visualize_sentence_token_attention_heatmap(
   attentions, tokenizer, input_ids,
   sentence_boundaries, sentences,
   sent_x_idx=0, sent_y_idx=1  # Select sentences
)
```

For token-level analysis, **SAVIS** integrates with BertViz (Vig, 2019). This integration allows users to combine our sentence-level aggregation with established token-level visualization approaches:

```
# Prepare attention data for BertViz
attention_data, tokens, sentence_b_start = \
   isa.get_sentence_token_attention(
       sentence_x_idx=0,
       sentence_y_idx=1
   )

# Visualize with BertViz
from bertviz import head_view
head_view(attention_data, tokens)
```

3.3 Interactive Exploration

SAVIS includes interactive features for exploring attention patterns across documents. Users can select specific layers and attention heads for detailed analysis, and switch between different visualization modes:

```
# Configure visualization parameters
vis.visualize_sentence_attention(
   figsize=(15,10)  # Adjust plot size
)

# Layer/head-specific analysis
vis.visualize_token_attention_heatmap(
   attentions, tokenizer, input_ids,
   layer=5,      # Specific layer
   head=3,       # Specific head
   figsize=(50,50)
)
```

Figure 2 shows how **SAVIS** visualizes attention structures. In this example, we compare standard prompting, which shows direct attention flows, with Chain-of-Thought (CoT; Wei et al., 2022) prompting, where our visualization highlights sequential attention patterns between intermediate reasoning steps. The figure demonstrates the application of these strategies on two types of NLP tasks: math word problems (Cobbe et al., 2021) and CommonsenseQA (CSQA; Talmor et al., 2019). Math word problems assess language models' arithmetic capabilities through mathematical scenarios. CSQA presents commonsense questions that often require prior knowledge. Standard prompting yields direct responses, while CoT prompting generates intermediate reasoning steps. The ISA patterns show that CoT prompting creates stronger connections between sentences, indicating a more structured reasoning process.

4 Case Studies

To evaluate **SAVIS**'s effectiveness in real-world scenarios, we conducted two case studies with public relations (PR) practitioners. Using the Gemma 7B model (Team et al., 2024), we analyzed how PR practitioners utilized our tool to improve their analysis of public communications. Table 1 summarizes the quantitative results.

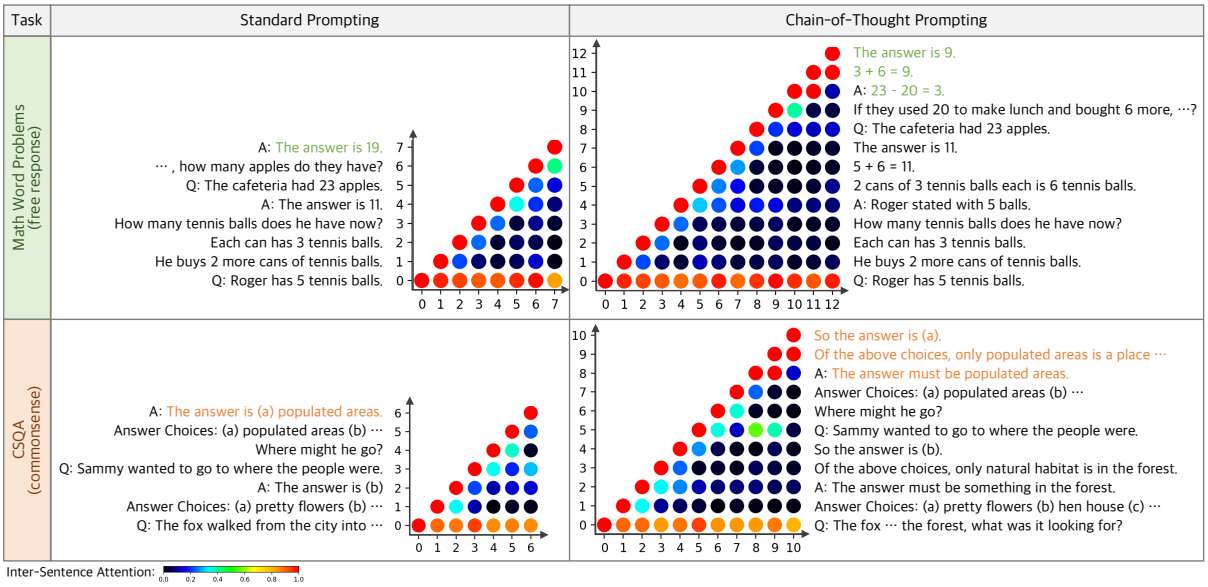


Figure 2: Comparison of ISA patterns between standard and CoT prompting. Standard prompting (left) directly answers questions without explicit reasoning steps. CoT prompting (right) offers answers through a logical progression of sentences based on intermediate reasoning steps to reach the final answer. The inter-sentence attention visualization demonstrates how each sentence in CoT prompting contributes to the understanding and generation of subsequent sentences. LLM-generated text is highlighted. Note that the first sentence starts from 0.

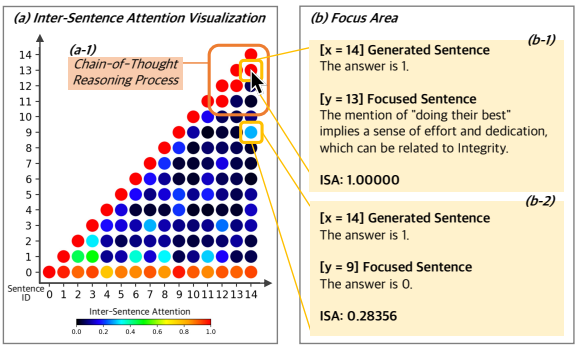


Figure 3: Interactive visualization on the analysis of reviews on government policies. (a) Inter-sentence attention visualization with ISA values. (a-1) CoT reasoning steps. (b) Sentence display on hover interaction. (b-1) High ISA between conclusion and rating. (b-2) Reference to similar example during generation.

These case studies were designed to evaluate the Organization-Public Relationship Assessment (OPRA; Kim and Ni, 2013) theory, a key framework in PR, using **SAVIS**. OPRA assesses the health and strength of relationships between organizations and the public across four dimensions: *trust*, *satisfaction*, *commitment*, and *control mutuality* (Lee and Jun, 2013; Liu and Ni, 2021).

4.1 Case Study 1: Analysis of Environmental Policy Public Comments

In the first case study, PR practitioners analyzed public comments regarding proposed changes to the National Environmental Policy Act (NEPA),

Table 1: Quantitative results of case studies.

Metric	NEPA	Amazon
Dataset Size (documents)	100	230
Initial Accuracy (%)		
Trust	78	82.61
Satisfaction	76	81.74
Commitment	79	83.48
Control Mutuality	77	82.17
Average	77.5	82.50
Final Accuracy (%)		
Trust	**93**	**89.13**
Satisfaction	**94**	**90.00**
Commitment	**95**	**90.87**
Control Mutuality	**94**	**89.57**
Average	**94**	**89.89**
ISA Threshold	0.10	0.25

collected from regulations.gov. The dataset comprised 100 public comments, evenly split between the Trump administration's NEPA revision period and the Biden administration's subsequent policy changes.

The practitioners first conducted an initial LLM analysis using standard prompts, achieving an average accuracy of 77.5% across all OPRA dimensions in sentiment classification. Using **SAVIS**, they identified a systematic bias in the LLM's interpretation of mixed-sentiment comments. For instance, when analyzing the comment *"I am writing in strong support of the National Environmental Policy Act because NEPA is the bedrock of our work to ensure full protection of important places for birds and people,"* the LLM initially classified trust as 0,

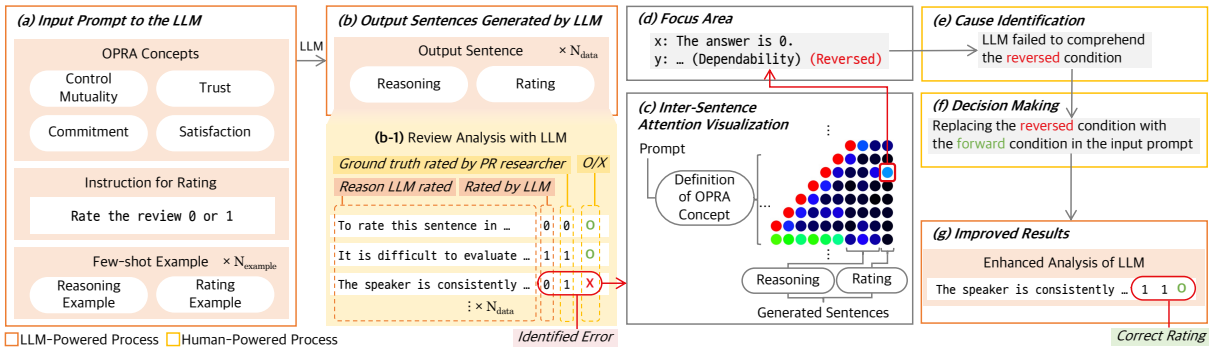


Figure 4: Analysis of Amazon product reviews. (a) Input prompt with OPRA concepts and instructions. (b) LLM output showing reasoning steps and ratings. (b-1) Discrepancy between LLM ratings and ground truth. (c) Inter-sentence attention visualization. (d) Focus area examination. (e) Error identification in reversed conditions. (f) Input prompt refinement. (g) Corrected ratings after refinement.

contradicting the explicit support for NEPA. The visualization revealed a low ISA value of 0.06 for this critical first sentence in the final rating step.

Based on these observations, the practitioners implemented a CoT process in their prompts. As shown in Figure 3 (a-1), this modification encouraged the LLM to generate intermediate reasoning steps before producing a final rating. When a practitioner hovers over a point of interest in Figure 3 (a), the generated and focused sentences are displayed as shown in Figure 3 (b). Figure 3 (b-1) reveals that the sentence at y=13, corresponding to the conclusion of the CoT, impacted the rating generated in sentence x=14 (ISA=1.0). Additionally, Figure 3 (b-2) indicates that the sentence at x=14 referenced the similar example sentence at y=9 during its generation (ISA=0.28356). Following this implementation, the results improved. As detailed in Table 1, NEPA's Final Accuracy shows consistently high performance across all OPRA dimensions: Trust (93%), Satisfaction (94%), Commitment (95%), and Control Mutuality (94%).

4.2 Case Study 2: Sentiment Analysis of E-commerce Customer Review

In the second case study, PR practitioners analyzed a dataset of 230 product reviews from Amazon. Figure 4 illustrates their comprehensive analysis process using **SAVIS**. The process began with an input prompt outlining OPRA concepts and instructions for rating (Figure 4 (a)). In Figure 4 (b), the LLM output consists of reasoning steps and final ratings. Through this structured output, practitioners identified discrepancies between the LLM's ratings and ground truth as shown in Figure 4 (b-1).

Through the inter-sentence attention visualization in Figure 4 (c) and a critical examination of the focus area in Figure 4 (d), the practitioners identified that the LLM struggled particularly with reviews containing reversed conditions. For example, in reviews stating *"I would have given 5 stars if..."*, the LLM misclassified these as positive reviews, achieving an initial average accuracy of 82.5%. The cause of this error was due to the LLM's misunderstanding of reversed conditions.

Based on these insights, the practitioners modified the input prompts by replacing reversed conditions with forward conditions, as shown in Figure 4 (f). This adjustment aimed to eliminate the LLM's confusion in interpreting conditional statements. Finally, as demonstrated in Figure 4 (g), this refinement successfully corrected the previously misclassified ratings. As shown in Table 1, the accuracy in Amazon review classification increased across all OPRA dimensions, ranging from 89.13% to 90.87%. This improvement highlighted the importance of understanding how LLMs process reversed conditions in reviews, consistent with recent studies (Jang et al., 2022; Truong et al., 2023).

5 Comparative Analysis

We evaluated **SAVIS** against **BertViz** (Vig, 2019), **Dodrio** (Wang et al., 2021), and a **baseline** approach without visualization. Twenty participants analyzed the NEPA public comments (n=100) and Amazon product reviews (n=230) for 60 minutes each. Table 2 and Figure 5 show the results.

Using **SAVIS**, participants identified more errors (NEPA: 7.2, Amazon: 7.8) than with other tools, except when using the combination of **SAVIS**+BertViz (NEPA: 8.8, Amazon: 9.0). Participants also identified errors faster with **SAVIS** (NEPA: 7.9 minutes, Amazon: 7.4 minutes) compared to BertViz (NEPA: 11.5 minutes, Amazon:

Table 2: Comparison of approaches with and without attention visualization for analyzing PR datasets. The numbers represent averages from 20 participants. Initial accuracies for the four OPRA dimensions are shown in Table 1.

Metric	**SAVIS** (Ours)		**BertViz** (Vig, 2019)		**SAVIS + BertViz**		**Dodrio** (Wang et al., 2021)		**Baseline** (w/o Visualization)	
Environment	Jupyter		Jupyter		Jupyter		Web		Jupyter	
Dataset	NEPA	Amazon	NEPA	Amazon	NEPA	Amazon	NEPA	Amazon	NEPA	Amazon
Number of Identified Errors	7.2	7.8	5.3	6.1	**8.8**	**9.0**	6.9	7.4	1.5	1.8
Time to Identify Error (min)	**7.9**	**7.4**	11.5	11.6	8.3	7.5	8.8	8.3	46.3	42.9
Number of Prompt Revisions	**2.3**	3.4	3.5	4.3	2.4	**3.3**	2.5	3.4	4.4	5.3
Final Accuracy (%)										
Trust	89.50	85.87	86.00	83.48	**91.50**	**87.83**	88.00	84.78	83.00	83.04
Satisfaction	88.00	84.78	86.50	83.70	**92.00**	**87.61**	87.50	84.35	83.50	83.26
Commitment	89.00	83.91	87.00	83.91	**92.50**	**88.04**	88.50	84.13	84.00	83.48
Control Mutuality	87.50	85.22	85.50	83.70	**91.00**	**87.39**	87.00	84.57	82.00	82.61
Average	88.50	84.95	86.25	83.70	**91.75**	**87.72**	87.75	84.46	83.13	83.10

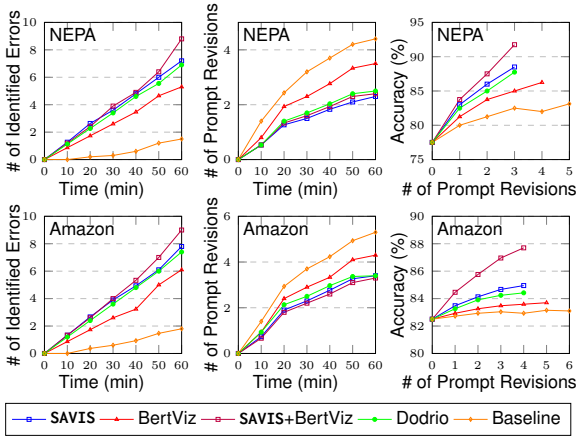


Figure 5: Comparative analysis of **SAVIS** against other visualization tools.

11.6 minutes) and other approaches. They needed fewer prompt revisions with **SAVIS** (NEPA: 2.3, Amazon: 3.4) and achieved higher accuracy. Starting from the initial values (NEPA: 77.5%, Amazon: 82.5%), participants using **SAVIS** improved accuracy to 88.5% (+11 percentage points (pp)) for NEPA and 84.95% (+2.45pp) for Amazon. Those using **SAVIS**+BertViz showed slightly higher improvements (NEPA: 91.75%, +14.25pp; Amazon: 87.72%, +5.22pp). Dodrio and BertViz also enabled improvements, while the baseline approach had minimal improvement for both datasets.

Figure 5 shows performance over time and across revisions. Participants using **SAVIS** and **SAVIS**+BertViz detected errors faster, needed fewer revisions, and reached peak accuracy in fewer iterations compared to other approaches.

6 Limitations and Future Work

While our visualization tool demonstrates effectiveness for PR tasks, several limitations should be acknowledged. Our approach has been primarily evaluated with a single model architecture, and testing across diverse architectures would be necessary to confirm broader applicability. While **SAVIS**'s sentence-level visualization effectively enables PR practitioners to identify patterns quickly, token-level visualizations can complement it for detailed analysis. Our comparative analysis shows that while **SAVIS** achieves strong performance independently, combining it with token-level visualizations like BertViz provided additional benefits, suggesting potential value in multi-level visualization approaches.

For future work, we plan to develop a unified interface that dynamically combines sentence-level and token-level views, allowing users to begin with simplified patterns and drill down to details where needed. We aim to explore adaptive visualization techniques that automatically suggest appropriate levels of detail for different analysis tasks.

7 Conclusion

This paper introduced **SAVIS**, a novel visualization tool for interpreting LLM text generation at the sentence level. We defined inter-sentence attention (ISA) to analyze sentence relationships and implemented a visualization approach combining interactive features and attention patterns. **SAVIS** is open-source and available at `https://pypi.org/project/savis` under MIT license.

Our case studies on NEPA public comments and Amazon product reviews demonstrated **SAVIS**'s effectiveness in identifying and addressing LLM output inaccuracies, particularly in handling complex linguistic structures such as reversed conditions. Comparative analysis validated our approach, showing that practitioners using **SAVIS** achieved faster error detection and higher accuracy with fewer revisions than existing tools. This approach offers practitioners an efficient tool for understanding LLM behavior in various text analysis tasks.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2024-00460980, Development of Federated Computing (deFAI, decentralized federated AI) for Enhanced Utilization of Distributed Data and its Application, 50%) and (No.2022-0-00305, Development of automatic data semantic information composition/expression technology based on augmented analysis for diagnosing industrial data status and maximizing improvement, 50%).

Ethical Statement

We adhere to the ethical principles outlined in the ACL Code of Ethics throughout our research. We considered potential ethical implications when creating and applying **SAVIS**. Our work aims to enhance the transparency and interpretability of LLMs, promoting responsible AI technology use through a better understanding of model behavior. All data used in our case studies and comparative analysis were meticulously reviewed to exclude personally identifiable information, ensuring privacy protection. We advocate for the ethical use of LLMs, encouraging practitioners to employ visualization tools in ways that are respectful, equitable, and conscious of societal impacts.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, and 109 others. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Adam Coscia and Alex Endert. 2023. Knowledgevis: Interpreting language models by comparing fill-in-the-blank prompts. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–13.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jeong-Nam Kim and Lan Ni. 2013. Two types of public relations problems and integrating formative and evaluative research: A review of research programs within the behavioral, strategic management paradigm. *Journal of Public Relations Research*, 25(1):1–29.

Hyung Min Lee and Jong Woo Jun. 2013. Explicating public diplomacy as organization–public relationship (opr): An empirical investigation of oprs between the us embassy in seoul and south korean college students. *Journal of Public Relations Research*, 25(5):411–425.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.

Shusen Liu, Tao Li, Zhimin Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. 2018. Visual interrogation of attention-based models for natural language inference and machine comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 36–41, Brussels, Belgium. Association for Computational Linguistics.

Wenlin Liu and Lan Ni. 2021. Relationship matters: How government organization-public relationship impacts disaster recovery outcomes among multiethnic communities. *Public Relations Review*, 47(3):102047.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Zijie J. Wang, Robert Turko, and Duen Horng Chau. 2021. Dodrio: Exploring transformer models with interactive visualization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 132–141, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).