Designing Writing Assistants for Scientific Figure Captions: A Thematic Analysis

Ho Yin (Sam) Ng[1], Ting-Yao Hsu[1], Jiyoo Min[2], Sungchul Kim[3], Ryan A. Rossi[3], Tong Yu[3], Hyunggu Jung[4], Ting-Hao 'Kenneth' Huang[1],

[1]Pennsylvania State University, [2]University of Seoul, [3]Adobe Research, [4]Seoul National University

Correspondence: sam.ng@psu.edu

Abstract

Scientific figure captions are essential for communicating complex data but are often overlooked, leading to unclear or redundant descriptions. While many studies focus on generating captions as an 'output', little attention has been given to the writer's process of crafting captions for scientific figures. This study examines how researchers use AI-generated captions to support caption writing. Through thematic analysis of interviews and video recordings with 18 participants from diverse disciplines, we identified four key themes: (1) integrating captions with figures and text, (2) bridging gaps between language proficiency and domain expertise, (3) leveraging multiple AI-generated suggestions, and (4) adapting to diverse writing norms. These findings provide actionable design insights for developing AI writing assistants that better support researchers in creating effective scientific figure captions.

1 Introduction and Backgrounds

Scientific figures communicate complex data and concepts to readers in research papers (Durbin Jr, 2004). These figures are accompanied by captions, providing essential context and explanations to enhance the reader's understanding of the presented information (Qian et al., 2021). Writing figure captions may seem straightforward, but many researchers overlook them, resulting in unclear explanations that confuse readers (Jambor et al., 2021; Huang et al., 2023). Crafting a good caption demands clarity, brevity, and alignment with the figure's purpose, making it more challenging than it appears. It requires specialized language and detailed explanations to effectively communicate abstract and complex scientific concepts (Gomez-Perez and Ortega, 2019). The difficulty of this task has contributed to the prevalence of low-quality captions in scientific literature (Huang et al., 2023), highlighting the need for improved approaches to caption writing.

Meanwhile, artificial intelligence (AI), especially large language models (LLMs), offers seemingly promising solutions for producing reasonable quality captions (Anagnostopoulou et al., 2024; Liew and Mueller, 2022; Rotstein et al., 2024; Gopu et al., 2023). For example, the SciCap project (Hsu et al., 2021) compiled a large dataset of scientific figures and captions from arXiv papers to develop models for generating high-quality captions for scientific figures. Many caption-generation models have been proposed for scientific figures (Rojas and Carranza, 2024; Cao and Liu, 2024; Singh et al., 2023; Wu et al., 2024). Despite these advancements, there remains a limited understanding of how AI-generated captions benefit writers of scholarly papers. While prior research has demonstrated that AI-generated captions are effective from a reader's perspective, as shown through human evaluation methods (Zhang et al., 2024; Aguirre et al., 2023; Hsu et al., 2023), their utility for writers has been underexplored from a Human-Computer Interaction (HCI) perspective. Prior studies often only focused on readers' perspectives—having people evaluate AI-generated captions by providing ratings or feedback—rather than examining the writing process itself from the writers' perspective. Recent efforts have started to address this gap. For instance, SciCapenter showed that AI-generated captions can reduce cognitive load for writers (Hsu et al., 2024), and another study investigated how different configurations and inputs improve caption generation to assist writers (Ng et al.). However, these efforts emphasize quantitative measures, such as cognitive load or usability of AI outputs, and fall short of capturing qualitative, higher-level insights from practitioners engaged in the caption-writing process, which can guide the design of future writing assistants.

This paper seeks to address this gap by examining how scholarly paper writers interact with

Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025), pages 1–10

AI-generated captions during the writing process through a *qualitative* lens. We analyzed video recordings and transcripts from a think-aloud study (Ng et al.) in which participants rewrote figure captions for their previously published papers, as well as their post-study interview responses. Using thematic analysis (Clarke and Braun, 2017), guided by a design space for writing assistants proposed by Lee *et al.* (Lee et al., 2024), we tailored the framework to the unique context of scientific figure captions. We identified four main themes in the study data: *(i)* the multimodal and complex context inherent in figure caption writing, *(ii)* the gaps between domain-specific knowledge and linguistic expression, especially in describing complex scientific concepts in English, *(iii)* the diverse ways participants utilized AI-generated suggestions, and *(iv)* the variations in norms and conventions for figure captions across different academic disciplines. By identifying these challenges and insights, this paper seeks to bridge the gap between current AI capabilities and the specific needs of scientific writers, contributing to the advancement of more effective and intuitive writing assistance technologies.

2 Methods

2.1 Data

We acquired the video recordings and transcripts collected in a prior study by Ng *et al.* (Ng et al.), which involved 18 participants from diverse research fields. The participant pool included researchers from Computer Science/Informatics (28%), Human-Computer Interaction (22%), Artificial Intelligence/Robotics (17%), and other fields such as Energy and Minerals Engineering, Mechanical Engineering, Environmental Engineering, Chemistry/Biochemistry, Materials Science, and Cybersecurity (6% each). Participants were aged 22 to 44, with the majority (78%) between 26 and 29 years old. 72% of participants reported that English was not their first language. We briefly outline their study protocol below.

Original Study: Caption Re-Writing Study and Interview. The original study used a mixed-methods approach that combined writing tasks, think-aloud protocols, and semi-structured interviews (Ng et al.). Sessions were conducted via Zoom and lasted approximately one hour. The procedure consists of three main steps: **(1) Pre-task Interview**, participants described their typical caption-writing process, figure creation methods, and characteristics of effective captions.

(2) Writing Task, participants received a Google Doc link to rewrite two captions from their previously published works. They were provided with three configurations of AI-generated captions using GPT-4o, which varied by input type and output length:

1. UNLIMITED: Figure image and reference paragraphs as input, with no output length restrictions.

2. 30-WORD: Same inputs as UNLIMITED, but output limited to 30 words.

3. TEXT-ONLY: Reference paragraphs only as input (no image), with unlimited output length.

An example of these caption generation configurations is provided in Appendix A (Fig. 2). Participants could use these AI-generated captions in any way they found helpful while completing their task. Throughout the process, participants verbalized their thoughts using a think-aloud protocol.

(3) Post-task Interview, participants reflected on the AI-generated options, suggested improvements for AI tools, and compared their rewritten captions to the originals.

2.2 Analysis Approach

We conducted a qualitative analysis to explore how participants interact with AI-generated captions, how they write captions, and how they view AI use for scientific figures.

We adopted an existing design space for intelligent writing assistants by Lee *et al.* (Lee et al., 2024) as our deductive framework, applying its five *aspects*, *i.e.*, Task, User, Technology, Interaction, and Ecosystem, to organize and interpret the data. We used thematic analysis (Clarke and Braun, 2017) on interview transcripts and video recordings as follows: First, the first author of this paper reviewed all the transcripts and manually annotated text spans relevant to the five aspects. Then, these annotation were then grouped into specific codes (Fig. 1). Finally, the codes were synthesized into higher-level themes, capturing areas of agreement and divergence among participants. As a result, we identified four main themes from the data, which we describe in the section below.

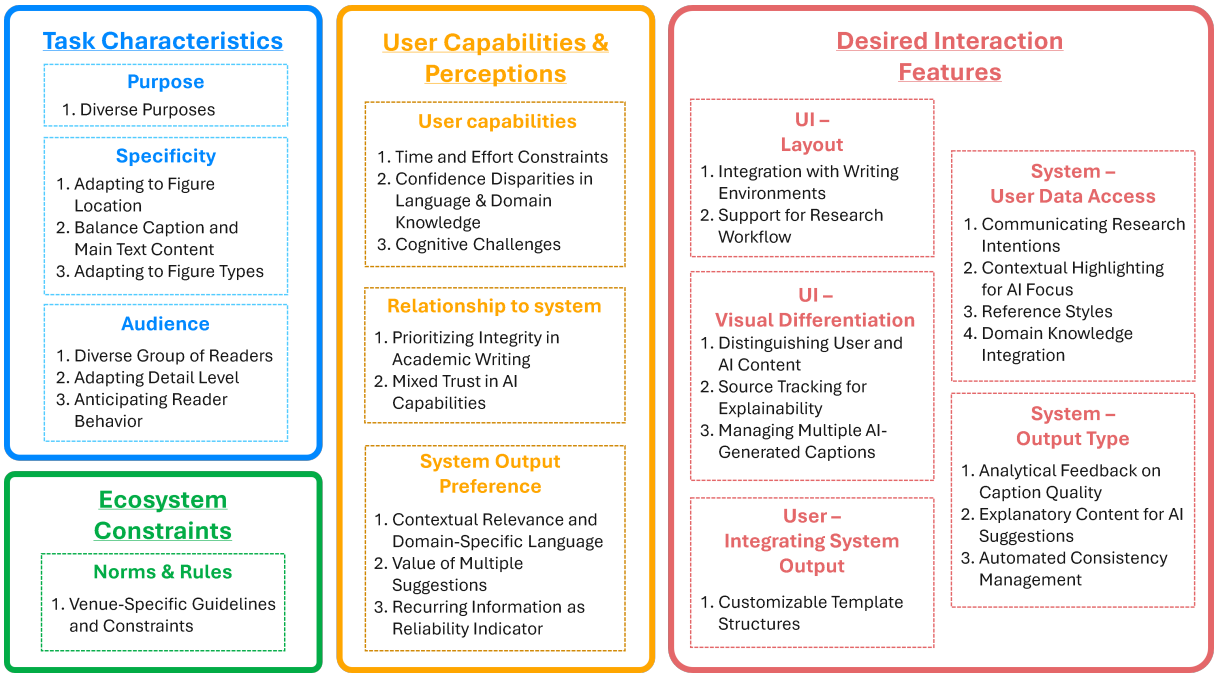


Figure 1: Codes developed for qualitative analysis of researchers' interactions with AI-generated captions, categorized under Lee *et al.* (Lee et al., 2024)'s design space aspects: TASK, USER, INTERACTION, and ECOSYSTEM.

3 Findings

Our analysis identified four key themes regarding researchers' experiences with AI-generated captions for scientific figure caption writing. These themes corresponded to the **TASK, USER, INTERACTION, AND ECOSYSTEM** aspects of the guiding design space (Lee et al., 2024), while the TECHNOLOGY aspect was less prominent in participants' concerns. Below, we detail the four themes that emerged from our thematic analysis, noting their alignment with the relevant design space aspects. In the following, we used the participant labels (P1, P2, P3, etc.) from the original interview study. Keeping these labels maintains consistency between our analysis and the prior work.

3.1 Considering Figures, Captions, and Text in a Multi-modal Context (TASK)

Our findings reveal a strong connection between figures, captions, and main text in research papers. Participants stressed the need for AI-generated captions to align with each figure and its context (P2, P3, P7, P8, P11, P12). For example, P2 highlighted the importance of context awareness: "*It depends on the figures in different sections of the research papers. If it is in the results section or the methods section, we will use the precise [caption].*" Moreover, our findings highlight the importance of treating captions, figures, and main text as an interconnected whole (P1, P2, P3, P7, P11, P15, P16). As P15 explained: "*It depends on the situation or context of the paragraph. Sometimes I write text first and then add the captions and images [figures]. But sometimes, if I already have images [figures], I make sentences around them.*"

Many participants also emphasized the importance of avoiding redundancy among figures, captions, and the main text (P7, P9, P11, P14, P16, P17). As P11 noted: "*Add details and data in captions or under figures that you didn't mention in the text of your paper... if you have described or mentioned some of the details that is [sic] visible in the figure, there is no need to repeat that information over and over again in the caption.*"

3.2 Gaps in Confidence Across Language and Domain Knowledge (USER)

Our findings indicate a clear contrast in participants' confidence regarding language proficiency versus technical or domain expertise. Many participants, especially non-native English speakers, reported lower confidence in writing captions due to language concerns (P4, P5, P8, P10, P15). As P8 noted: "*Difficult to write these long captions because for me it's hard to construct nice and beautiful sentences.*" By contrast, participants generally felt more assured in their domain knowledge (P1, P11, P15, P16, P18). Several even believed their

expertise surpassed the capabilities of AI. For example, P1 remarked: "*I feel like I know best. And so I would do what I think is best. I feel like I might know better than AI on how to caption the figure on my paper.*"

This theme highlights a key challenge in caption writing: many researchers, especially non-native English speakers, struggle with language despite their technical and domain expertise. For captions specifically, this language barrier is significant because captions must clearly communicate complex visual information independently.

AI captioning tools can address this gap by complementing researchers' domain knowledge with language support: Researchers verify scientific accuracy, while AI improves linguistic clarity. This collaboration directly addresses the unique demands of figure captions, helping researchers create clear, accessible visual explanations for diverse scientific audiences.

3.3 Leveraging Multiple (AI) Suggestions and Perspectives (INTERACTION)

In the original study, participants were presented with multiple AI suggestions generated by different approaches. Our analysis shows that offering multiple perspectives—despite being generated by AI instead of humans—can be beneficial, as it could inspire paper writers. Paper writers can explore different angles instead of relying on a single solution, thereby enhancing creativity and decision-making. We further break it down into two types of usages:

3.3.1 Inspiration Through Diversity of Suggestions

Participants valued AI's ability to present multiple approaches to caption writing, often using these ideas as inspiration rather than direct answers. Many participants (P2, P3, P4, P6, P9, P10, P11, P13, P14, P15, P16, P17, P18) incorporated multiple suggestions into their work, finding it helpful to compare options and select the most useful elements for their final captions. P3 offered an insightful analogy: "*[It] feels like having three extra collaborators write captions for me and then I'm like cherry picking different parts to write my caption.*" This collaborative view highlights how AI can supplement, rather than replace, human creativity in scientific writing.

3.3.2 Trust Through Repetition of Suggestions

A notable finding emerged on how participants handled multiple AI-generated suggestions. Participants often used a comparative approach, trusting elements that appeared consistently across different outputs. As P10 noted: "*I will read all the suggestions and think about what is [sic] the common things in the captions, so which means that kind of information is important.*" Several participants (P6, P10, P14) observed that seeing similar content across AI suggestions influenced their own writing, guiding them to adopt particular phrasings or details. This observation reveals a potential cognitive bias in AI writing assistants, where repetition across suggestions may inadvertently shape researchers' perceptions of what is important or accurate. Recognizing this effect is essential for designing AI tools that support, rather than unduly influence, scientific communication.

3.4 Adapting to Diverse Norms in Scientific Writing (ECOSYSTEM)

Our analysis identified significant variations in caption writing practices across scientific disciplines and publication venues, shaped by explicit venue-specific requirements and implicit discipline-specific styles. It highlights the complex challenges researchers face when crafting captions. These challenges involve balancing formal guidelines with unwritten conventions:

3.4.1 Explicit Venue-Specific Requirements

Participants stressed the importance of following explicit guidelines set by conferences and journals, highlighting a need for flexible AI writing assistants. Several participants (P4, P6, P10, P11, P12, P18) noted challenges related to page or word limits and specific formatting rules. As P10 explained: "*A lot of conference and journal have different limits. Sometimes I want to write more information, but I have to cut down some of it.*" This tension between providing comprehensive captions and adhering to publication constraints suggests that AI tools should be capable of tailoring output to specific venue requirements, such as word count or formatting rules.

3.4.2 Implicit Discipline-Specific Styles

Beyond explicit guidelines, variations in caption styles across disciplines presented a more implicit challenge. Many participants (P2, P5, P7, P8, P9, P11, P12, P15) reported relying on examples from

their field to guide their caption writing. As P15 described: "*If I make the captions for the [figure], then first I refer to other papers because there are a lot of papers about with the same or similar topics.*". This reliance on field-specific examples highlights the influence of unwritten disciplinary norms on caption writing. These norms are often understood within the community but not explicitly documented. Some participants also noted highly specific conventions unique to their fields. For instance, P9 remarked: "*I don't know other majors and other research papers, how they arrange their papers. But I think for data science area, it is not professional to include numbers [in captions].*"

These findings reveal a wide range of implicit writing styles across disciplines that researchers learn through exposure and practice rather than formal guidelines.

4 Discussion

Our analysis identifies four key themes that can guide the development of more effective writing assistants for scientific figure captions: (1) integrating captions with figures and text, (2) addressing gaps between language proficiency and domain expertise, (3) utilizing multiple AI-generated suggestions, and (4) accommodating diverse writing norms. In this section, we propose practical design recommendations for future caption-writing tools, using two illustrative examples to highlight strengths and limitations: SciCapenter (Hsu et al., 2024), which generates captions with quality ratings and contextual information to aid refinement (see Appendix B, Fig. 3), and FigurA11y (Singh et al., 2024), which focuses on accessibility by creating alt text drafts and offering interactive revision tools (see Appendix C, Fig. 4). By analyzing these systems, we identify gaps in current approaches and offer insights to guide the development of more versatile and user-centered AI writing assistants.

4.1 Design Suggestions

4.1.1 Integrating Captions with Figure and Text

Our study showed that writers often struggle to maintain consistency between captions, figures, and main text. While SciCapenter partially addresses this need by displaying related figure-mentioning paragraphs alongside captions, providing useful context during caption editing. However, it lacks deeper integration between captions and the broader manuscript structure for the writer to tracing the connection easily.

Recommendation. Future AI caption writing tools could enable interactive linking between captions, figures, and text to improve consistency and reduce redundancy:

1. **Interactive Linking and Visualization**: Create clickable, color-coded links between figure components, captions, and related text sections, allowing researchers to easily trace relationships between different elements of their manuscript, enhancing overall coherence.

2. **Automated Consistency Checking**: Implement automated checks to flag discrepancies in terminology or data representations, prompting researchers to review and refine content for improved accuracy and coherence throughout their manuscripts.

4.1.2 Bridging Language Gaps While Incorporating Domain Expertise

AI tools excel at generating linguistically coherent captions but often struggle with nuanced domain-specific knowledge. While systems like FigurA11y demonstrate the potential of human-AI collaboration, they still have limitations in understanding complex domain-specific relationships.

Recommendation. Future AI caption writing tools could combine AI language capabilities with user domain expertise:

1. **Domain Knowledge Input Interface**: Allow researchers to input key domain concepts or terminology, guiding AI outputs to ensure captions are tailored to specific disciplines or venues. This could involve developing an interface where users can upload custom glossaries or select from a searchable ontology of domain-specific terms, which would help the AI model generate more accurate and relevant captions.

2. **AI Confidence Highlighting and Output Refinement**: Develop AI models that assess their confidence in generated content, highlighting areas of low confidence for user refinement, thus leveraging human expertise to ensure scientific accuracy.

4.1.3 Leveraging Multiple AI Suggestions

Our study revealed that diverse AI suggestions inspire creativity. SciCapenter generates multiple options with quality ratings, but lacks diversity in focusing on different aspects of the figure (e.g., methods vs. results).

Recommendation. Future AI caption writing tools could generate and combine diverse suggestions:

1. **Multi-prompt Generation**: Implement parallel prompting strategies using different instruction sets (e.g., focusing on visual elements, data relationships, or research implications).

2. **Interactive combination interface**: Provide a modular editing environment where users can combine elements from multiple suggestions, such as drag-and-drop paragraph components.

4.1.4 Adapting to Diverse Writing Norms

Participants noted that caption styles vary across disciplines and venues. While existing systems like SciCapenter provide general-purpose solutions, they lack customization for specific norms. For example, it does not allow users to tailor captions to discipline-specific styles or venue requirements.

Recommendation. Future AI caption writing tools could adapt to different writing contexts:

1. **Venue-Specific Template**: Offer preconfigured templates based on common guidelines from major journals to ensure compliance with submission standards (e.g. word limits, formatting conventions).

2. **Exemplar-Based Learning**: Analyze captions from similar publications within a discipline to generate outputs aligned with established norms, using visually or contextually similar figures as guides.

4.2 Limitations

Our study provides valuable insights into how researchers interact with AI-generated captions for scientific figures, but it has limitations that should be addressed in future research. First, the original study's controlled environment, where participants rewrote captions for their previously published papers, may not fully capture the complexities of real-world scientific writing scenarios. Typically, paper authors write captions for works in progress rather than published papers, which presents different challenges and considerations. Second, while we refer to SciCapenter and FigurA11y as examples to illustrate design suggestions, these systems differ significantly from the original study setup. In the study, participants received AI-generated captions through Google Docs in a one-way interaction—they could not prompt the AI for refinements or engage in iterative feedback. This contrasts with SciCapenter and FigurA11y, which offer interactive caption refinement capabilities. Our work provides foundational insights into researchers' needs that could enhance these and future systems.

5 Conclusion and Future Work

This study explored how researchers interact with AI-generated captions to improve scientific figure caption writing. By conducting thematic analysis of interviews and video recordings, we identified four key themes: (1) integrating captions with figures and text, (2) bridging gaps between language proficiency and domain expertise, (3) leveraging multiple AI-generated suggestions, and (4) adapting to diverse writing norms. These themes highlight the unique challenges of caption writing and provide actionable insights for designing AI writing assistants. By focusing on the writer's process rather than just the output, this research contributes to a deeper understanding of how AI can assist researchers in crafting effective figure captions. These insights lay the groundwork for developing more effective and intuitive AI tools that enhance scientific communication.

Building on these insights, future research should focus on developing and testing AI tools for scientific caption writing in real-world scenarios. Such evaluations will reveal their effectiveness and usability while providing deeper insights into researchers' needs and challenges. Observations of authentic writing practices will guide refinements, ensuring that AI systems address the complexities of caption writing across disciplines. This work will lead to more adaptable, user-centered AI solutions that enhance both the writing process and the quality of scientific communication.

References

Carlos Aguirre, Shiye Cao, Amama Mahmood, and Chien-Ming Huang. 2023. Crowdsourcing thumbnail

captions: Data collection and validation. *ACM Transactions on Interactive Intelligent Systems*, 13(3):1–28.

Aliki Anagnostopoulou, Thiago Gouvea, and Daniel Sonntag. 2024. Enhancing journalism with ai: A study of contextualized image captioning for news articles using llms and lmms. *arXiv preprint arXiv:2408.04331*.

Stanley Cao and Kevin Liu. 2024. Figuring out figures: Using textual references to caption scientific figures. *arXiv preprint arXiv:2407.11008*.

Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology*, 12(3):297–298.

Charles G Durbin Jr. 2004. Effective use of tables and figures in abstracts, presentations, and papers. *Respiratory care*, 49(10):1233–1237.

Jose Manuel Gomez-Perez and Raul Ortega. 2019. Look, read and enrich-learning from scientific figures and their captions. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 101–108.

Arunkumar Gopu, Pratyush Nishchal, Vishesh Mittal, and Kuna Srinidhi. 2023. Image captioning using deep learning techniques. In *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, volume 1, pages 1–5. IEEE.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. Scicap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264.

Ting-Yao Hsu, Chieh-Yang Huang, Shih-Hong Huang, Ryan Rossi, Sungchul Kim, Tong Yu, C Lee Giles, and Ting-Hao Kenneth Huang. 2024. Scicapenter: Supporting caption composition for scientific figures with machine-generated captions and ratings. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Ting-Yao Hsu, Chieh-Yang Huang, Ryan Rossi, Sungchul Kim, C Lee Giles, and Ting-Hao K Huang. 2023. Gpt-4 as an effective zero-shot evaluator for scientific figure captions. *arXiv preprint arXiv:2310.15405*.

Chieh-Yang Huang, Ting-Yao Hsu, Ryan Rossi, Ani Nenkova, Sungchul Kim, Gromit Yeuk-Yin Chan, Eunyee Koh, Clyde Lee Giles, and Ting-Hao'Kenneth' Huang. 2023. Summaries as captions: Generating figure captions for scientific documents with automated text summarization. *arXiv preprint arXiv:2302.12324*.

Helena Jambor, Alberto Antonietti, Bradly Alicea, Tracy L Audisio, Susann Auer, Vivek Bhardwaj, Steven J Burgess, Iuliia Ferling, Małgorzata Anna Gazda, Luke H Hoeppner, et al. 2021. Creating clear and informative image-based figures for scientific publications. *PLoS biology*, 19(3):e3001161.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–35.

Ashley Liew and Klaus Mueller. 2022. Using large language models to generate engaging captions for data visualizations. *arXiv preprint arXiv:2212.14047*.

Ho Yin Sam Ng, Ting-Yao Hsu, Jiyoo Min, Sungchul Kim, Ryan A Rossi, Tong Yu, Hyunggu Jung, and Ting-Hao Kenneth Huang. Understanding how paper writers use ai-generated captions in figure caption writing. In *2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle*.

Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, Joel Chan, Ryan A Rossi, Sana Malik, and Tak Yeon Lee. 2021. Generating accurate caption units for figure captioning. In *Proceedings of the Web Conference 2021*, pages 2792–2804.

Mateo Alejandro Rojas and Rafael Carranza. 2024. Enhancing scientific figure captioning through cross-modal learning. *arXiv preprint arXiv:2406.17047*.

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5689–5700.

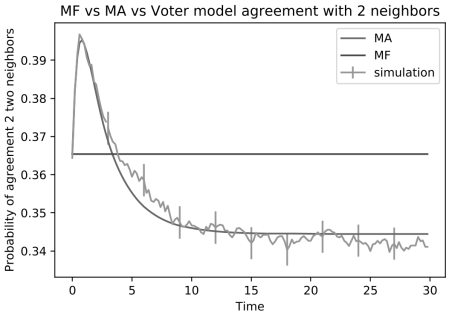
Ashish Singh, Prateek Agarwal, Zixuan Huang, Arpita Singh, Tong Yu, Sungchul Kim, Victor Bursztyn, Nikos Vlassis, and Ryan A Rossi. 2023. Figcaps-hf: A figure-to-caption generative framework and benchmark with human feedback. *arXiv preprint arXiv:2307.10867*.

Nikhil Singh, Lucy Lu Wang, and Jonathan Bragg. 2024. Figura11y: Ai assistance for writing scientific alt text. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 886–906.

Jian Wu, Börje F Karlsson, and Manabu Okumura. 2024. Caption alignment and structure-aware attention for scientific table-to-text generation. *IEEE Access*.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, et al. 2024. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *arXiv preprint arXiv:2406.10522*.

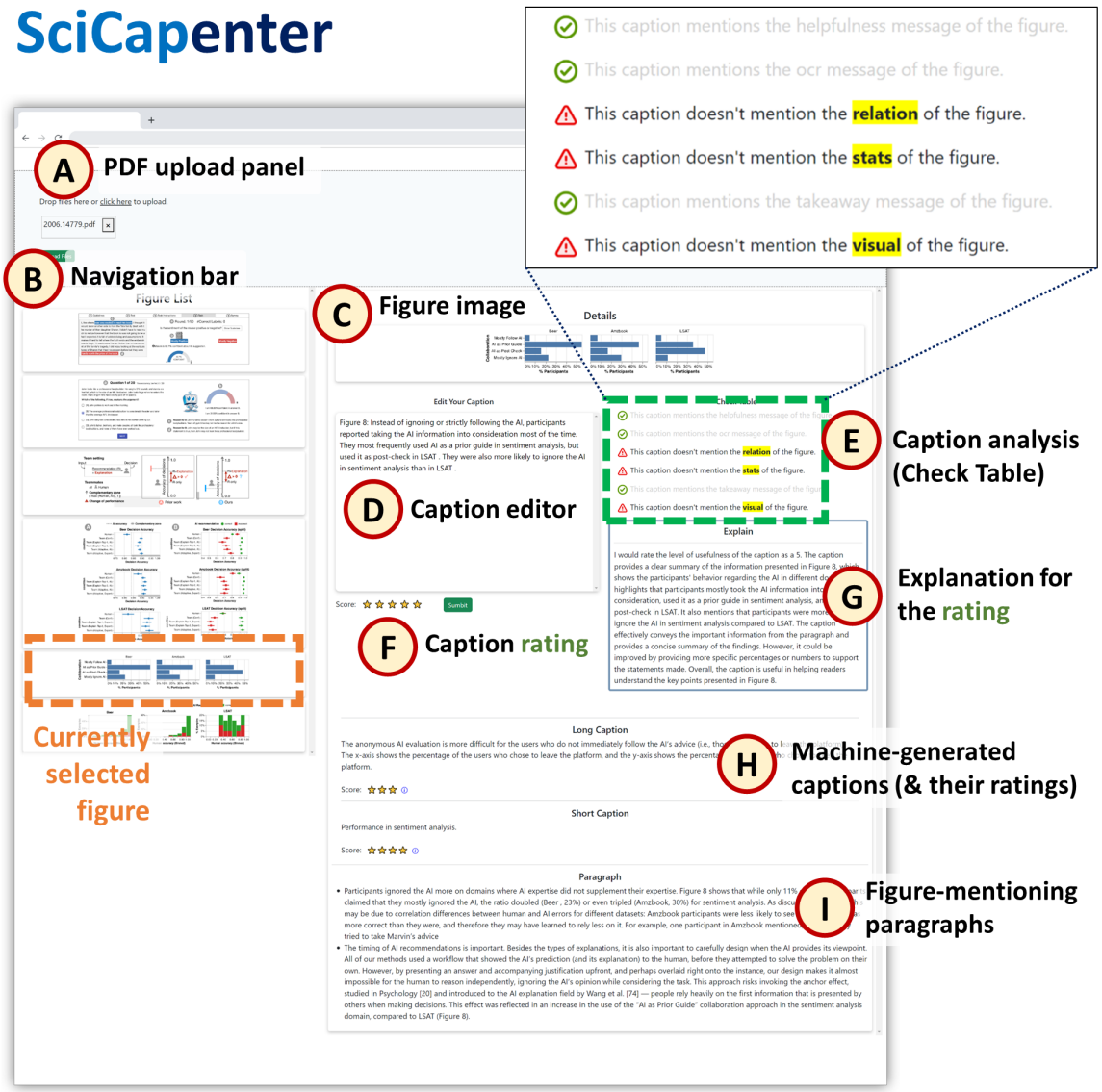
A AI-Generated Caption Example



Model	Generated Caption
IMG+TEXT (30-WORD)	Comparison of mean-field (MF) and Markovian approximation (MA) models with simulation for voter model agreement probability on a 3-regular tree, showing improved accuracy of the Markovian approach.
TEXT-ONLY (UNLIMITED)	Comparison of the ad hoc Markovianized local equation and the original local equation for the voter model on a 3-regular tree truncated after 9 generations, demonstrating improved accuracy in predicting the probability of root agreement with two neighbors.
IMG+TEXT (UNLIMITED)	Comparison of Mean-Field (MF), Markovian Approximation (MA), and simulation results for the probability of agreement with two neighbors in a voter model on a 3-regular tree.

Figure 2: GPT-4o generated captions in three different configurations: (i) GPT-4o (image+text) with a 30-word limit, (ii) GPT-4o (text-only) with unlimited length, and (iii) GPT-4o (image+text) with unlimited length. Reprinted From (Ng et al.).

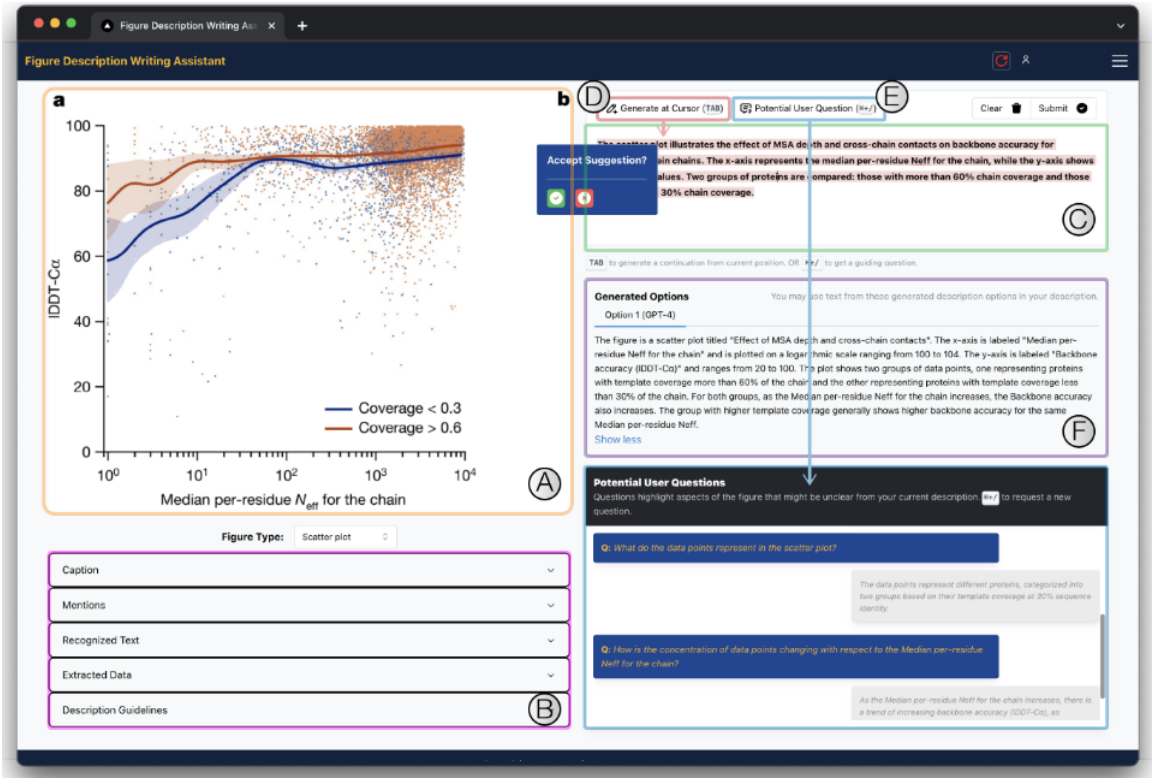
B Design of SciCapenter



[This image shows the interface of the SciCapenter system, which includes several key components for document and caption management.]

Figure 3: Overview of SciCapenter system interface. **PDF Upload Panel (A):** A drag-and-drop interface for uploading PDF files. **Navigation Bar (B):** A horizontal bar showing a list of figures extracted from the uploaded document. **Figure Image (C):** The main area displaying the image of the selected figure. **Caption Editor (D):** A text box for editing the caption of the selected figure. **Caption Rating (F):** A feedback system that allows GPT to rate the quality of the caption, represented by a star rating. **Caption Analysis (Check Table) (E):** Icons indicating the presence or absence of key elements in the caption, such as helpfulness or takeaway message. **Explanation for the Rating (G):** A textual explanation providing insight into why a particular star rating was given to the caption. **Machine-generated Captions & Their Ratings (H):** This section includes long and short captions generated by AI models, each accompanied by their respective star ratings. **Figure-mentioning Paragraphs (I):** Paragraphs in the document that mention the target figure, providing context or additional information.

C Design of FIGURA11Y



[This image shows the interface of the FIGURA11Y system, which includes several key components for document and caption management.]

Figure 4: Overview of FIGURA11Y system interface. On the left, it shows (A) the figure and (B) extracted metadata. On the right, it shows (C) the description authoring field, (D) the *Generate at Cursor* feature with generated initial text below, (E) the *Potential User Questions* request button and results, and (F) a pre-generated draft description.