

Five Years of SCICAP: What We Learned and Future Directions for Scientific Figure Captioning

Ting-Hao ‘Kenneth’ Huang¹, Ryan A. Rossi², Sungchul Kim², Tong Yu²,
Ting-Yao ‘Edward’ Hsu¹, Ho Yin Ng (Sam)¹, Clyde Lee Giles¹

¹The Pennsylvania State University, 201 Old Main, University Park, PA, USA

²Adobe Research, San Jose, CA, USA

{txh710, txh357, sam.ng, clg20}@psu.edu

{ryrossi, sukim, tyu}@adobe.com

Abstract

Between 2021 and 2025, the SCICAP project grew from a small seed-funded idea at The Pennsylvania State University (Penn State) into one of the central efforts shaping the scientific figure-captioning landscape. Supported by a Penn State seed grant, Adobe, and the Alfred P. Sloan Foundation, what began as our attempt to test whether domain-specific training, which was successful in text models like SciBERT, could also work for figure captions expanded into a multi-institution collaboration. Over these five years, we curated, released, and continually updated a large collection of figure-caption pairs from arXiv papers, conducted extensive automatic and human evaluations on both generated and author-written captions, navigated the rapid rise of large language models (LLMs), launched annual challenges, and built interactive systems that help scientists write better captions. In this piece, we look back at the first five years of SCICAP and summarize the key technical and methodological lessons we learned. We then outline five major unsolved challenges and propose directions for the next phase of research in scientific figure captioning.

2021-2025: The First Five Years of SCICAP

A scholar once shared a piece of wisdom: a good researcher either “makes something big out of something small” or “makes something small out of something big.” In other words, one can choose a narrow, well-scoped question and explore it deeply, or take on an ambitious, moonshot idea by starting from its simplest form. The first five years of SCICAP belong to the former path. We chose a laser-focused, narrowly-scoped problem—**how to generate and evaluate scientific figure captions**—and pursued it with persistence and depth. What followed was a journey that began with a small idea and grew into a sustained, multi-institution collaboration that continues today.

This section reviews this journey from 2021 to 2025.

Origin Story of SCICAP: 133k+ Figure-Caption Pairs from arXiv Papers

The story of SCICAP began in late 2019, when we at Penn State started discussing whether we could build a large-scale dataset of real-world scientific figures and captions to support multimodal modeling for scientific visuals. The spark

came when Lee walked into Kenneth’s office and asked, “*Can we build a CaptionBERT?*” Lee knew Kenneth had worked on vision-and-language datasets and models before, including the Visual Storytelling (VIST) dataset, and around that time domain-specific BERT variants, such as SciBERT (Beltagy, Lo, and Cohan 2019) and BioBERT (Lee et al. 2019), were showing clear advantages. Their success made us wonder whether scientific figure captions might also benefit from domain-specific modeling, if only we could gather enough real data.

We soon drafted a seed-grant proposal to Penn State outlining two motivations: First, figure captions in scientific papers were often poorly written, frequently lacking the detail or clarity needed for readers to understand a figure without returning to the full text. This seemed like an ideal place for automated assistance. Second, if we wanted to build technology to improve caption quality, there was a fundamental data gap: no large-scale, real-world dataset of scientific figures paired with their captions existed. Most prior datasets were either synthetic or too small to support reliable captioning models (Kafle et al. 2018; Kahou et al. 2017; Chen et al. 2019, 2020). We aimed to fix both problems at once.

The proposal was funded. By 2021, we had constructed and released the first version of the SCICAP dataset (Hsu, Giles, and Huang 2021), focusing on **graph plots from computer-science and machine-learning arXiv papers**. We gathered 295,028 papers published between 2010 and 2020, extracted over 2.17 million figures, identified 416,804 graph plots (19.2%), and after filtering out compound figures, produced a clean set of **133,543 single-panel figures**.

In the 2021 SCICAP release paper (Hsu, Giles, and Huang 2021), we also evaluated several baseline vision-to-language models. Their performance was strikingly poor: captions were often low-quality, inaccurate, or even outright gibberish. The results confirmed our intuition: scientific figure captioning was far from a solved problem, and real progress would require both better data and better models.

A Highly Contextual Task: Captioning as Contextual Summary, Not Just Visual Description

Soon after we released the first version of SCICAP, we learned that Adobe Research had also been exploring scientific figure captioning (Qian et al. 2021). We reached out, and after a few very encouraging conversations, our two

groups decided to join forces in Fall 2021. That collaboration ended up shaping the direction of SCICAP for years.

As we examined the low-quality outputs from our initial models, it became clear that scientific figure captioning is fundamentally a context-rich problem. In one of our early meetings, an Adobe collaborator asked a question that stuck with us: “Where does the information in a caption actually come from?” No one, not even a human expert, can write a good caption from the image alone.

From 2021 to early 2022—before ChatGPT or multi-modal LLMs were publicly available—we worked together to figure out how to incorporate document context, especially the paragraphs that mention each figure (e.g., “Figure 3 shows...”), into the captioning process. However, building effective multimodal models proved difficult, and early attempts to combine images with text yielded only modest improvements.

Gradually, we noticed a pattern: many scientific figure captions were essentially paraphrases or compressed summaries of the paper’s own descriptions. That observation led to a simple but surprising experiment. We treated figure captioning as a text-summarization problem and fine-tuned a summarization model, PEGASUS (Zhang et al. 2020), using only the figure-mentioning paragraphs—no images at all. The text-only approach worked remarkably well. We also found that roughly 75% of caption words could be traced directly to those paragraphs or the figure’s OCR.

This work resulted in our INLG 2023 paper, which received the Best Long Paper award and was selected as one of the three finalists for the Best Evaluation Paper award (Huang et al. 2023). It demonstrated that scientific figure captioning behaves much more like contextual summarization than visual description. Furthermore, a small human evaluation in the paper confirmed our motivating observation: many author-written captions were mediocre at best. PhD students rated over 50% of author-written captions as *unhelpful*. This result highlights the limitations of caption authoring practices and the potential for LLM assistance to improve caption utility for readers.

LLMs Judge Caption Quality Like PhD Students, Not Undergraduates

Then came ChatGPT in late 2022, a history-defining moment. For the first time, fluent, high-quality text generation became accessible to millions. After some quick empirical testing, we made a strategic decision: instead of building new architectures specifically for scientific figure captioning, we would focus on understanding and leveraging LLMs’ ability to generate captions for scientific figures. Our goal is to help authors and readers work with clearer, more useful captions, and it became clear that building on LLMs would be a more effective and generalizable path than creating application-specific models.

Once we reached some major milestones in caption generation, the next obvious challenge was evaluation. Our INLG paper had already reminded us of a familiar truth echoed across many prior text-generation projects: automatic evaluation is not enough. Traditional reference-based metrics like BLEU and ROUGE did not correlate reliably with human



Figure 1: The SCICAP logo, designed in 2023 for the first SCICAP Challenge and used across all related projects since.

judgments. And human evaluation (at least for scientific figure captions) was expensive and logistically difficult. Scientific captions require expert background knowledge; generic online crowd workers cannot evaluate captions for physics or electrical-engineering figures, and even NLP researchers are often unqualified outside their own subfields.

Against this backdrop, as well as with ChatGPT reshaping the field, we began experimenting with LLMs as automatic judges in early 2023. PhD students and trained undergraduates rated captions generated by different approaches, and we compared their ratings to those produced by LLMs. The results, published at EMNLP 2023 (Hsu et al. 2023), were interesting: LLM judgments aligned well with PhD students’ ratings. Undergraduate ratings, however, diverged. Less-experienced readers cared more about clear take-away messages, while experts prioritized numerical precision and technical detail.

Our study landed at the early wave of the “LLM-as-judge” trend and drew more attention than we expected. Interestingly, a later 2024 study would report that LLM judges sometimes favor their own generations (Panickssery, Bowman, and Feng 2025). Our work avoided that bias by deliberately excluding LLM-generated captions from the evaluation pool.

SCICAP Challenge: Building a Community and Expanding to 476k+ Figure-Caption Pairs

We launched the first SCICAP Challenge in 2023 (Figure 1) at the ICCV workshop Closing the Loop Between Vision and Language. We invited teams around the world to generate captions for an expanded version of our dataset, *i.e.*, the SCICAP Challenge dataset, which contained **476,389 single-panel scientific figures across 8 scholarly domains and 5 figure types**.¹ Each figure came with its image, OCR text, figure type, and the paragraphs in the arXiv that explicitly mentioned it. When we designed the specification of the SCICAP Challenge, we intentionally made the figure-mentioning paragraphs part of the standard input that ev-

¹The SCICAP Challenge dataset (Hsu et al. 2025) included figures from arXiv papers across computer science, physics, math, economics, statistics, quantitative biology, quantitative finance, and electrical engineering. Figures were categorized into node diagrams, equations, graph plots, scatterplots, and bar charts. We limited the per-paper sampling to encourage participation from teams with modest computational budgets. This dataset was more than three times the size of our original 2021 SCICAP release.

ery team received. We wanted this design choice to send a clear signal: scientific figure captioning is a highly context-dependent generation task; a good caption cannot be written from an image alone.

Built on the outputs submitted to the 2023 Challenge, we conducted a large-scale human and automatic evaluation study, later published at TACL and presented at EMNLP 2025 (Hsu et al. 2025). In this paper, we found that multimodal LLMs (MLLMs), particularly GPT-4V, consistently produced captions preferred by professional editors over all other approaches, including human-written captions. This was true across domains, across figure types, and even for papers published after the models’ claimed training cut-off date. In cases where authors wrote captions carelessly, MLLM-generated captions were often clearly better—a humbling result, but also an exciting one for the potential of assistive tools.

Since 2023, we have continued to run the SCICAP Challenge annually—at IJCAI 2024 and at the LM4Sci Workshop at COLM 2025. Over time, this effort has helped form a small yet active community working on scientific figure captioning. Some teams have even returned to participate again after their first participation; we even collaborated with a team to publish their approach (Kim et al. 2025). Our very first SCICAP paper (Hsu, Giles, and Huang 2021), which originally was a Findings paper without fancy modeling, has been cited over 130 times.

Constant Trade-offs Between Automatic and Human Evaluation. Each year we ran the SCICAP Challenge, we faced the same dilemma: automatic evaluation is cheap and fast but unreliable, while human evaluation is reliable but extremely slow and expensive. Across the 2023-2025 challenges, we experimented with different ways to balance this trade-off. In 2023, we gave two awards: one based on ROUGE-2-Normalized scores on the full hidden test set, and another on a curated subset of higher-quality captions, motivated by our finding that models performing well overall could still fail on well-written captions (Huang et al. 2023). However, rankings barely changed across the two sets. In 2024, we switched to human evaluation on a small sample of 200 figures and split the challenge into two tracks: a short-caption track, where systems had to produce captions no longer than the originals for at least 30% of samples, and a long-caption track with the opposite constraint. This time, the results diverged: teams outperformed the original authors of captions in the long-caption track, but authors still performed better in the short-caption track. This could be due to humans’ known preference for longer captions. In 2025, to mitigate this bias, we went with two distinct tracks, one judged by humans and one judged automatically, using the new LAMP-CAP dataset (Ng et al. 2025b) focused on personalized captioning. This time, the results diverged again across two sets. Teams were able to mimic the author’s style per automatic evaluations, but this often hurt caption quality perceived by humans. In fact, human evaluators consistently preferred GPT-4o’s vanilla captions over personalized ones.

If these three years taught us anything, it is that we still understand very little about how to reliably measure the

quality of scientific figure captions *at scale*. This particular type of captioning is highly contextual, often requires specialized knowledge to assess, and even author-written captions are not always reliable.

Supporting Caption Writers: Promising, but Highly Nuanced

As our data, generation methods, and evaluation practices began to mature, we turned to the part of the loop that matters most for real impact: helping writers. We focused on the writer side because readers have little control over caption quality—most readers cannot revise captions or run AI models—but improving the writing process can immediately benefit thousands of readers at once.

Our first step was building SCICAPENTER, an interactive caption-writing assistant that displays the figure, its figure-mentioning paragraphs, and several AI-generated draft captions, and allows writers to iteratively revise them with quality ratings and contextual guidance (Hsu et al. 2024). A pilot study suggested that these suggestions reduced cognitive load and helped writers produce higher-quality text. We then conducted a separate, deeper study in which 18 researchers rewrote captions from their own recently published papers while using multiple LLM-generated captions as starting points (Ng et al. 2025a). Through video-based interaction analysis, we found that writers typically began by copying or adapting AI-generated captions, preferred longer and detail-rich drafts, and relied heavily on the textual–visual alignment of the suggestions. However, they struggled when figures were conceptually complex or when AI suggestions conflicted with disciplinary norms. We also saw broader themes (Ng et al. 2025c): caption writing is deeply multimodal, writers often lack linguistic confidence but trust their domain expertise, they benefit from multiple diverse AI suggestions, and caption norms vary dramatically across fields.

Personalized Captioning Through Multimodal Profiles

In our caption-writing studies (Ng et al. 2025a,c), we noticed a recurring gap: even when AI-generated captions were of fine quality, authors often felt they were too generic and did not match their writing style or the stylistic conventions of their target venues. This pushed us to think seriously about personalization. Although personalized text generation is now common with LLMs (Zhang et al. 2024), the key practical question for our task was: *what information can we use to characterize the style of a group of co-authors?* Using all papers written by each co-author was an obvious answer, but it was expensive and conceptually messy, since scholarly writing is collaborative and we rarely know which author actually wrote which caption. So we turned to something more grounded and reliable: information within the same paper.

This led us to create LAMP-CAP (Ng et al. 2025b), a dataset derived from the SCICAP Challenge dataset and designed specifically for personalized scientific figure captioning using multimodal profiles. For each target figure, LAMP-CAP provides the figure image and its figure-mentioning paragraphs as model input, along with up to

three additional figures from the same paper (each with its own image, caption, and paragraphs) as a style profile. Across 110,828 target figures, we tested four LLMs and consistently found that adding these multimodal profiles helped models generate captions closer to the original author-written ones. Human evaluation further confirmed the usefulness of personalization. Interestingly, human judges preferred captions generated with one profile figure over both no-profile and all-profile conditions. This echoes a common dilemma in personalization (Zhang et al. 2024): a little tailoring improves quality and feels helpful to users, but pushing too hard to mimic the user’s style can hurt the generation quality.

Challenges and Future Directions

In this position paper, we argue that scientific figure captioning is a clear example of how a technology that appears powerful—LLM-generated captions for scientific figures—still faces many obstacles before it can be truly useful in real-world settings. Working on this problem has shown us how much depth and complexity can surface when we study real tasks that look small and well defined from the outside. It also shows a major opportunity for the technical HCI community, because many of the hardest remaining challenges in scientific figure captioning are fundamentally *human-centered*.

In the following, we describe the challenges and potential future directions that emerged from our five years of experience working on the SCICAP project and related efforts. These ideas grew out of ongoing discussions and continuous reflection within our team, and we believe sharing them can benefit the broader research community.

Challenge #1: Captioning with Incomplete or Missing Context

Real-world writing rarely unfolds in neat, fully drafted documents. To help authors, captioning systems must handle partial or missing context for figures in unfinished papers or for standalone visuals circulating on social media, in news, or in advertisements. **18.81% of figures in the original SCICAP dataset lacked any identified mentions** (Huang et al. 2023). Generating meaningful captions when only the image is available remains one of the field’s most practical yet unsolved problems.

Challenge #2: Captions for a Variety of Readers

SCICAP’s evaluations showed that reader expertise shapes caption preference (Hsu et al. 2023). Undergraduates valued clear takeaways; PhD students cared about methodological precision and numbers. Future models might adapt captions for different audiences, such as experts, students, and practitioners, which will require advances in audience modeling, controllable generation, and implicit preference inference. Building audience-aware captioning systems without relying on extensive user profiling remains an open challenge in personalization.

Challenge #3: Personalizing to the Writer’s Style

The experimental results from both the LAMP-CAP dataset (Ng et al. 2025b) and the SCICAP Challenge 2025 showed a clear trade-off: strong stylistic imitation can reduce factual accuracy or clarity. This tension between stylistic faithfulness and communicative quality raises important questions about authorship, identity, and the design of human-AI co-writing systems. It also brings up a broader question that some NLP researchers have asked: to support writing, do we actually need personalized text generation, or is generic generation sufficient for most authors to build on?

Challenge #4: Studying and Supporting the Writing Process

Most caption research evaluates outputs from the reader’s perspective, not the writer’s. Yet to understand how AI truly helps scientists, we must study caption writing in realistic conditions. Recruiting authors actively drafting their own papers is difficult; proxy studies, *e.g.*, rewriting old captions or captioning others’ figures (Ng et al. 2025a,c), only approximate the real task. The Sloan-funded deployment of SCICAPENTER opens a rare opportunity to observe genuine writing-in-progress (Huang and Fetzer 2024), but conducting such ecologically valid studies remains expensive and methodologically hard.

Challenge #5: Bridging Captioning and Figure Understanding

Despite shared goals, figure captioning and figure understanding research have mostly evolved in isolation. Work on figure parsing, chart reasoning, and figure QA rarely intersects with user-facing captioning tasks that need evaluation beyond parsing accuracy. Conversely, many captioning projects, including our own, did not fully evaluate faithfulness or correctness with respect to the visual content. Captioning focuses on communication for human readers, while figure understanding is often developed as a reasoning benchmark. From a captioning perspective, our approach to the task treated it as writing support, which allowed us to worry less about strict grounding because authors remained in the loop. Still, there are clear scenarios where visual faithfulness matters. Bringing these threads together could yield more holistic systems that both understand and explain scientific visuals, reduce duplication of effort, and push the field toward genuine multimodal comprehension.

Conclusion

Scientific figure captioning is a small, well-scoped task that turned out to be far more complex than we expected. Working on it taught us practical lessons about text generation, evaluation, and how AI-generated text can fit into real writing workflows. There is still much to explore, and we see this work as the starting point for the next phase of research on captioning and multimodal language technologies.

Acknowledgments

We thank all co-authors and collaborators on SCICAP-related projects, as well as every team that participated in

the SCICAP Challenges. This work has been supported by a Penn State IST Seed Grant (2020), Adobe Research’s gift (2022), and the Alfred P. Sloan Foundation grant (2024–2026, #G-2024-22721).

References

- Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pre-trained Language Model for Scientific Text. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. Hong Kong, China: Association for Computational Linguistics.
- Chen, C.; Zhang, R.; Kim, S.; Cohen, S.; Yu, T.; Rossi, R.; and Bunesco, R. 2019. Neural caption generation over figures. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC ’19 Adjunct, 482–485. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368698.
- Chen, C.; Zhang, R.; Koh, E.; Kim, S.; Cohen, S.; and Rossi, R. 2020. Figure Captioning with Relation Maps for Reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- Hsu, T.-Y.; Giles, C. L.; and Huang, T.-H. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3258–3264.
- Hsu, T.-Y.; Huang, C.-Y.; Huang, S.-H.; Rossi, R.; Kim, S.; Yu, T.; Giles, C. L.; and Huang, T.-H. K. 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*, CHI EA ’24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703317.
- Hsu, T.-Y.; Huang, C.-Y.; Rossi, R.; Kim, S.; Giles, C.; and Huang, T.-H. 2023. GPT-4 as an Effective Zero-Shot Evaluator for Scientific Figure Captions. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5464–5474. Singapore: Association for Computational Linguistics.
- Hsu, T.-Y. E.; Hsu, Y.-L.; Rohatgi, S.; Huang, C.-Y.; Ng, H. Y. S.; Rossi, R.; Kim, S.; Yu, T.; Ku, L.-W.; Giles, C. L.; et al. 2025. Do Large Multimodal Models Solve Caption Generation for Scientific Figures? Lessons Learned from SCICAP Challenge 2023. *arXiv preprint arXiv:2501.19353*.
- Huang, C.-Y.; Hsu, T.-Y.; Rossi, R.; Nenkova, A.; Kim, S.; Chan, G. Y.-Y.; Koh, E.; Giles, C. L.; and Huang, T.-H. 2023. Summaries as Captions: Generating Figure Captions for Scientific Documents with Automated Text Summarization. In Keet, C. M.; Lee, H.-Y.; and Zarriß, S., eds., *Proceedings of the 16th International Natural Language Generation Conference*, 80–92. Prague, Czechia: Association for Computational Linguistics.
- Huang, T.-H. K.; and Fetzer, M. 2024. IST researcher receives grant to improve scientific captions. *Penn State University News*. Last updated December 3, 2024.
- Kafle, K.; Price, B.; Cohen, S.; and Kanan, C. 2018. DVQA: Understanding Data Visualizations via Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kahou, S. E.; Michalski, V.; Atkinson, A.; Kádár, Á.; Trischler, A.; and Bengio, Y. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.
- Kim, J.; Lee, J.; Choi, H.-J.; Hsu, T.-Y.; Huang, C.-Y.; Kim, S.; Rossi, R.; Yu, T.; Giles, C. L.; Huang, T.-H. K.; and Choi, S. 2025. Multi-LLM Collaborative Caption Generation in Scientific Documents. In Wang, Q.; Yin, W.; Aich, A.; Suh, Y.; and Peng, K.-C., eds., *AI for Research and Scalable, Efficient Systems*, 142–160. Singapore: Springer Nature Singapore. ISBN 978-981-96-8912-5.
- Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.
- Ng, H. Y.; Hsu, T.-Y.; Min, J.; Kim, S.; Rossi, R. A.; Yu, T.; Jung, H.; and Huang, T.-H. K. 2025a. Understanding How Paper Writers Use AI-Generated Captions in Figure Caption Writing. In Wang, Q.; Yin, W.; Aich, A.; Suh, Y.; and Peng, K.-C., eds., *AI for Research and Scalable, Efficient Systems*, 173–192. Singapore: Springer Nature Singapore. ISBN 978-981-96-8912-5.
- Ng, H. Y. S.; Hsu, E.; Anantha Ramakrishnan, A.; Kveton, B.; Lipka, N.; DERNONCOURT, F.; Lee, D.; Yu, T.; Kim, S.; Rossi, R. A.; and Huang, T.-H. K. 2025b. LaMP-Cap: Personalized Figure Caption Generation With Multimodal Figure Profiles. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 9818–9832. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.
- Ng, H. Y. S.; Hsu, T.-Y.; Min, J.; Kim, S.; Rossi, R. A.; Yu, T.; Jung, H.; and Huang, T.-H. K. 2025c. Understanding Writing Assistants for Scientific Figure Captions: A Thematic Analysis. In Padmakumar, V.; Gero, K.; Wambsganss, T.; Serman, S.; Huang, T.-H.; Zhou, D.; and Chung, J., eds., *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, 1–10. Albuquerque, New Mexico, US: Association for Computational Linguistics. ISBN 979-8-89176-239-8.
- Panickssery, A.; Bowman, S. R.; and Feng, S. 2025. LLM evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.
- Qian, X.; Koh, E.; Du, F.; Kim, S.; Chan, J.; Rossi, R. A.; Malik, S.; and Lee, T. Y. 2021. Generating Accurate Caption Units for Figure Captioning. In *Proceedings of the Web Conference 2021, WWW ’21*, 2792–2804. New York,

NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, 11328–11339. PMLR.

Zhang, Z.; Rossi, R. A.; Kveton, B.; Shao, Y.; Yang, D.; Zamani, H.; Derroncourt, F.; Barrow, J.; Yu, T.; Kim, S.; et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.