

Collaborative Annotation as an Interface Metaphor for Personalizable AI Reading Support

Sireesh Gururaja
sgururaj@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Emma Strubell
strubell@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Abstract

In this position paper, we argue that collaborative annotation is a compelling interface metaphor through which to empower end-users to develop personalized AI reading support in the form of in-document highlights. Collaborative annotation confers many of the same benefits of traditional data annotation tools for specifying model behavior while not removing users from the context of the original document. This paradigm preserves the benefits of existing inline document markup tools, as well as the contextual, interpretive character of reading, while incorporating the flexibility and power of LLM-based annotation. Through what we call *collaborative machine annotation*, users can iteratively develop models as they refine their own mental models, and use the deep understanding of that model’s behavior for wider corpus analysis.

CCS Concepts

• **Human-centered computing** → **Collaborative interaction; Web-based interaction**; • **Computing methodologies** → **Information extraction**.

Keywords

Collaborative Annotation, Personalized Reading Support, Information Extraction

1 Introduction

Reading assistance tools have long made use of AI methods to provide in-context augmentation of document content. Tools like Scim [9] and PaperPlain [3] make use of machine learning to augment paper content to demonstrate a paper’s rhetorical structure, direct a reader’s focus, provide definitions, and more. These types of tools that provide in-context augmentation of paper contents are typically tightly scoped to an intended context of use and audience, and are therefore able to make assumptions about those contexts to deliver specific and reliable information; because the information is in-context on the document, it is also typically easily verifiable. However, assumptions about usage context and audience come at the cost of generality, even in the domains in which the tools are deployed. Scim highlights, for instance, generalize poorly outside of the experimental sciences, even as the Semantic Reader[18], in which Scim is deployed, makes them available on a much wider range of academic papers. Further, because these tools are based on models that rely on static schemas, they cannot account for readers’ preferences and idiosyncrasies in reading support; readers often rely on individually specific cues to make sense of papers [12].

By contrast, tools based on large language models (LLMs), whose behavior can be specified by users at inference time has led to the

proliferation of systems like SemanticScholar’s “ask this PDF” feature, or modes in popular LLM-based chatbots that allow uploaded documents that allow users to “chat” with the contents of a document. These systems are extremely flexible, and allow users to express complex and idiosyncratic information needs in natural language, and will sometimes anchor their responses to specific locations within documents. However, these chat-based tools introduce their own issues. While they allow end-user behavior specification and therefore personalized responses, verifying the reliability of the generated responses can be challenging. LLMs often exhibit a jagged capability frontier, where estimating where a model might succeed and fail is not always intuitive, and validating generated text is significantly more difficult to verify than in-context highlights. Additionally, these tools allow the specification of their behavior only in text, and keep no record of user preferences or reading focuses in between documents.

This paper argues that collaborative annotation—as exemplified by tools like Hypothesis [1] and Margin [2]—can provide an intuitive interface metaphor that allows for the flexibility of dynamic, user-specified LLM-based reading assistance, while emphasizing the user agency in evaluation, verifiability and context-awareness that inline highlighting tools can provide. Collaborative annotation between humans, in addition to its studied benefits for document comprehension, can provide in-context views of data to be annotated for granular specification of model behavior, and treating model suggestions as another collaborator can reduce the overhead for evaluating model performance and improving it with techniques like active learning. In-context experience with model suggestions also allows for deep, qualitative understanding of model performance such that users can make informed choices about model utility, differential use of model functions, and deploying the model to wider contexts, such as corpus-level annotation.

2 Collaborative Annotation as an Interface Metaphor

Collaborative annotation is a highly overloaded term in computing. It can refer to a variety of different processes, from the reconciliation of annotations in interfaces purpose-built for annotation such as Inception [17] or BRAT [24], to collaborative markup of educational materials as part of classroom assignments [20, 27]. In this paper, we focus primarily on the latter kind of collaborative annotation, and specifically tools like Hypothesis that enable collaborative annotation on any web content. We argue that this focus on in-context annotation allows for workflows similar to those in dedicated annotation tools, while presenting significantly less overhead than standalone annotation tools for web data.

Web-based collaborative annotation tools like Hypothesis make several design choices that create the advantages that we discuss later in this section. First, collaborative annotation tools make their annotations visible on the original document, whether webpage or PDF. In this way, collaborative annotation tools present an interface that is compatible with the inline highlighting and context-in-modal-dialog presentations of tools like Scim [9] and CiteSee [7]. Secondly, collaborative annotation tools already provide interface cues that indicate the authorship of annotations. Margin, for instance, provides a popup of the annotation author’s profile picture when a user mouses over their annotation; Hypothesis provides author metadata in their side panel’s annotations view. This signaling could be extended to indicate authorship, or even other factors, like confidence, with color, as in the case of Scim.

Because collaborative annotation tools already allow for the visualization of multiple annotators’ annotations in context, we can cast machine annotation as another participant in a collaborative annotation process, whose annotations can be viewed as suggestions for users to accept or reject, while maintaining their separation from a user’s own annotations. This of course does not preclude multiple human participants, who can make choices about how to view each other’s annotations, in addition to the machine’s, up to and including collaborative creation of annotation schemes. The visualization of machine annotation additionally allows users the same verification and qualitative understanding of a model’s behavior as existing inline annotation tools. We term this paradigm—where machines are engaged as a collaborative annotator—*collaborative machine annotation*

2.1 Behavior Specification, Evaluation, and Refinement

While prompting is a powerful paradigm for eliciting behavior from language models, prompting alone can sometimes be insufficient for highly specific tasks. Fonseca and Cohen [10] and Halterman and Keith [13], for example, find that while LLMs demonstrate some ability to follow annotation guidelines, their zero-shot performance is far from perfect; Halterman and Keith [13] explicitly suggests that fine-tuning can make models more effective at following codebooks, which requires some degree of data annotation.

Collaborative annotation tools, as the name suggests, provide an apt way to annotate this data, especially in the low volumes required for techniques like in-context learning [6] or LoRA fine-tuning [14]. While they do not have the fine-grained control and reporting of purpose-built annotation tools like Inception [17] do, they nonetheless provide affordances for multiple annotators to view each other’s annotations, reconcile them, and then potentially use them for machine learning training. Further, collaborative annotation tools can effectively leverage the context of the document content, which has been shown to affect labeling in subjective tasks like toxic speech detection [21].

Because of the separability of annotators, the visualization of annotations in the collaborative paradigm is also amenable to techniques like active learning, in which a model might suggest candidate instances that would benefit its performance, allowing rapid iteration and performance improvement of models without users having to leave the collaborative annotation interface. Models across

learning iterations can also be visualized as separate annotators, allowing users to compare the output of successive iterations on the same document, bringing more immediate interactivity to the process of model development. The in-context evaluation and qualitative understanding of a model’s performance also allows a user to understand the model’s performance in fine-grained ways, supporting wider, corpus-level annotation with some degree of reliability. This aligns with findings in Shankar et al. [22], a project which implemented a tool for using LLMs to perform corpus-level semantic processing; participants in the study found that verifying the quality of LLM-based annotation was a major hurdle in the adoption of pipelines involving LLMs.

2.2 Iterative, Adversarial Refinement of Mental Models

Reading support with fixed schemas, as in existing tools, often resembles information extraction. However, personalization depends on the user developing a framework for that extraction, a process much more similar to the process of developing codes for qualitative coding. As discovered in work investigating AI support for qualitative coding, a challenge to AI as a collaborative annotator is the degree to which AI can support the work not only of annotating documents, but also the interpretive work of developing a schema with which to annotate. Jiang et al. [15], for instance, find that HCI researchers engaged in qualitative coding wish to retain their agency in the process of developing codes, while [19] find that researchers want annotation to assist them with coding, once a code book has been developed. Collaborative machine annotation fits well within this framework: machine annotations can be constrained to only be applied based on codes already developed by human annotators.

In this paradigm, however, as the models that annotate document content are iteratively developed, from initial prompt onward, they can also serve to provide an adversarial view of the codes developed by researchers. In effect, while the models do not perfectly capture the human annotator’s intent, they may nonetheless provide instructive examples of alternate interpretations of a user’s provided prompt and examples. This might allow the user to consider the boundaries of their proposed codes and iterate on them, similar to the process of "active annotation" proposed in Vlachos [26]. This process might lead to users adding, splitting, or refining codes. However, noise must be a primary concern in the design of interfaces—while some degree of error may be generative, interfaces should allow users to tune the confidence of models such that the user is not confined to correcting bad model predictions.

2.3 Preservation of Serendipity and Mastery

A key consideration in our proposal for collaborative machine annotation is the preservation of serendipity in the process of reading documents. Serendipity, which is often studied in the contexts of information retrieval and recommender systems [11, 16, 25], is conceptualized in Binst et al. [5] as “*a user experience in which the user unintentionally encounters content that feels fortuitous, refreshing, and enriching.*” We argue that in keeping the user in the context of the original document, collaborative annotation allows for the discovery of content that users were not already searching for. By

contrast, in an LLM summary, a user is often provided with abstracted summaries, and pointed to the originating content only to verify the content of the summary.

Further, as argued in both [4, 23], naive use of AI tools can negatively impact skill formation. Bastani et al. [4] investigate this in a classroom mathematics setting, finding that without guardrails, use of generative AI assistance can negatively impact learners performance when access to the AI is lost; Shen and Tamkin [23] investigate learning in the context of code, finding similarly. In both cases, however, generative AI could be used in ways that did not cause this learning loss. Whether collaborative machine annotation can function similarly—whether the contextual nature, developing and iterating on personal reading assistance—should be investigated, especially in light of evidence of collaborative annotation’s positive impact in educational settings [8, 27, inter alia]

3 Conclusion

In this paper, we argue for *collaborative machine annotation*, i.e. the use of collaborative annotation as an interaction metaphor for end-user developed, personalized AI reading support. We argue that collaborative annotation can serve the purpose of annotating data for evaluating and developing models that perform in-context highlighting in documents, while at the same time allowing iterative refinement of a user’s mental models and preserving the contextual, interpretive nature of reading.

References

- [1] [n. d.]. Collaborate & Annotate with Hypothesis | Online Annotation Tool. <https://web.hypothes.is/>
- [2] [n. d.]. Margin. <https://margin.at/home>
- [3] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* 30, 5 (Sept. 2023), 74:1–74:38. doi:10.1145/3589955
- [4] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakçı, and Rei Mariman. 2025. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences* 122, 26 (July 2025), e2422633122. doi:10.1073/pnas.2422633122
- [5] Brett Binst, Lien Michiels, and Annelien Smets. 2025. What Is Serendipity? An Interview Study to Conceptualize Experienced Serendipity in Recommender Systems. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*. Association for Computing Machinery, New York, NY, USA, 243–252. doi:10.1145/3699682.3728325
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- [7] Joseph Chee Chang, Amy X. Zhang, Jonathan Bragg, Andrew Head, Kyle Lo, Doug Downey, and Daniel S. Weld. 2023. CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3544548.3580847
- [8] Jeffrey Clapp, Matthew DeCoursey, Sze Wah Sarah Lee, and Kris Li. 2021. “Something fruitful for all of us”: Social annotation as a signature pedagogy for literature education. *Arts and Humanities in Higher Education* 20, 3 (July 2021), 295–319. doi:10.1177/1474022220915128
- [9] Raymond Fok, Hita Kambhmettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 476–490. doi:10.1145/3581641.3584034
- [10] Marcio Fonseca and Shay Cohen. 2024. Can Large Language Models Follow Concept Annotation Guidelines? A Case Study on Scientific and Financial Domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8027–8042. doi:10.18653/v1/2024.findings-acl.478
- [11] Zhe Fu, Xi Niu, Xiangcheng Wu, and Ruhani Rahman. 2025. A Deep Learning Model for Cross-Domain Serendipity Recommendations. *ACM Trans. Recomm. Syst.* 3, 3 (March 2025), 29:1–29:21. doi:10.1145/3690654
- [12] Sireesh Gururaja, Nupoor Gandhi, Jeremiah Milbauer, and Emma Strubell. 2025. Beyond Text: Characterizing Domain Expert Needs in Document Research. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 4732–4745. doi:10.18653/v1/2025.findings-acl.244
- [13] Andrew Halterman and Katherine A. Keith. 2025. Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts. *Political Analysis* (Sept. 2025), 1–17. doi:10.1017/pan.2025.10017
- [14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. doi:10.48550/arXiv.2106.09685 arXiv:2106.09685 [cs].
- [15] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 94:1–94:23. doi:10.1145/3449168
- [16] Li Kang, Yuhuan Zhao, and Li Chen. 2025. Exploring the Potential of LLMs for Serendipity Evaluation in Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. Association for Computing Machinery, New York, NY, USA, 746–754. doi:10.1145/3705328.3748167
- [17] Jan-Christoph Klie, Michael Bugert, Beto Boulosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (Santa Fe, USA). Association for Computational Linguistics, 5–9. <http://tubiblio.ulb.tu-darmstadt.de/106270/> Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).
- [18] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu B. Kang, Egor Klevak, Bailey Kuehl, Michael J. Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2024. The Semantic Reader Project. *Commun. ACM* 67, 10 (Sept. 2024), 50–61. doi:10.1145/3659096
- [19] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173922
- [20] Kelly Miller, Sacha Zyto, David Karger, Junehee Yoo, and Eric Mazur. 2016. Analysis of student engagement in an online annotation system in the context of a flipped introductory physics class. *Physical Review Physics Education Research* 12, 2 (Dec. 2016), 020143. doi:10.1103/PhysRevPhysEducRes.12.020143
- [21] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 4296–4305. doi:10.18653/v1/2020.acl-main.396 Conference Name: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- [22] Shreya Shankar, Bhavya Chopra, Mawil Hasan, Stephen Lee, Björn Hartmann, Joseph M. Hellerstein, Aditya G. Parameswaran, and Eugene Wu. 2025. Steering Semantic Data Processing With DocWrangler. doi:10.48550/arXiv.2504.14764 arXiv:2504.14764 [cs].
- [23] Judy Hanwen Shen and Alex Tamkin. 2026. How AI Impacts Skill Formation. doi:10.48550/arXiv.2601.20245 arXiv:2601.20245 [cs].
- [24] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. Association for Computational Linguistics, Avignon, France.
- [25] Yu Tokutake, Kazushi Okamoto, Kei Harada, Atsushi Shibata, and Koki Karube. 2025. A Universal Framework for Offline Serendipity Evaluation in Recommender Systems via Large Language Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*. Association

- for Computing Machinery, New York, NY, USA, 5294–5298. doi:10.1145/3746252.3760911
- [26] Andreas Vlachos. 2006. Active Annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*. <https://aclanthology.org/W06-2209/>
- [27] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful classroom deployment of a social document annotation system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1883–1892. doi:10.1145/2207676.2208326

Received 12 February 2026