# ATLAS: A System for PDF-centric Human Interaction Data Collection

**Alexa Siu**[1], **Zichao Wang**[1], **Joshua Hoeflich**[3], **Naman Kapasi**[2]
**Ani Nenkova**[1], **Tong Sun**[1]

[1]Adobe Research [2]University of Calibornia, Berkeley [3]Northwestern University

{asiu,jackwa,nenkova,tsun}@adobe.com

## Abstract

The Portable Document Format (PDF) is a popular format for distributing digital documents. Datasets on PDF reading behaviors and interactions remain limited due to the challenges of instrumenting PDF readers for these data collection tasks. We present ATLAS, a data collection tool designed to better support researchers in collecting rich PDF-centric datasets from users. ATLAS supports researchers in programmatically creating a user interface for data collection that is ready to share with annotators. It includes a toolkit and an extensible schema to easily customize the data collection tasks for a variety of purposes, allowing the collection of PDF annotations (e.g., highlights, drawings) as well as reading behavior analytics (e.g., page scroll, text selections). We open-source ATLAS[1] to support future research efforts and review use cases of ATLAS that showcase our system's broad applicability.

## 1 Introduction

Collecting high-quality datasets from humans across varied domains has been one of the core driving factors for advances in artificial intelligence (AI) (Zha et al., 2023; Shneiderman, 2022). Recent progress in AI only makes the importance of such data more pronounced; as a canonical example, methods such as reinforcement learning from human feedback (RLHF) (Schulman et al., 2017; Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022), one of the core technology for large language model (LLM) fine-tuning and alignment (Liu et al., 2023b; Wang et al., 2023), critically depends on large-scale, high-quality human data such as chat, preference ranking, and question answering for both model training and evaluation (Zhao et al., 2024; Hendrycks et al., 2021; Talmor et al., 2019; Ni et al., 2019). In addition, the capabilities of AI systems are highly context-dependent and subjectively interpreted depending on the context being used and users' backgrounds (Denton et al., 2021; Lee et al., 2022). Thus, datasets within varied interaction contexts and collected from diverse users can help address these challenges by revealing the possibilities and limitations of AI systems. For example, datasets are central for evaluating Language Model's (LMs) capabilities in Natural Language Processing (NLP) research (Gehrmann et al., 2023). In Human-Computer Interaction (HCI) research, interaction datasets can help designers understand the capabilities of AI technology and inform interaction design and user experience design choices (Lee et al., 2022; Cuadra et al., 2021; Theodorou et al., 2021).

One of the most critical ingredients to human data collection is a tool that supports such deeds. The past few years have seen a surging need to collect such data, stimulating the rapid development of data collection tools and systems that cover various tasks and diverse modalities.[2] However, currently and notably missing from the landscape are tools that *support human data collection on digital documents in the form of the Portable Document Format (PDF)*. Indeed, the PDF is one of the most popular digital document formats with an estimated 2.5 trillion PDFs in the world today (Still, 2020). It is extensively employed across various industries such as healthcare, government, education, finance, legal, and e-commerce as the *de facto* standard for transactions, documentation, and communication. However, few, if any, datasets exist that comprehensively capture users' interactions with PDF documents. We posit one reason for this scarcity of PDF datasets is the challenging nature of process-

[1]https://github.com/frictionlessweb/documentstudies/

[2]A few examples include https://labelbox.com/, https://labelstud.io/, and https://appen.com/



*Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 87–96
June 16-21, 2024 

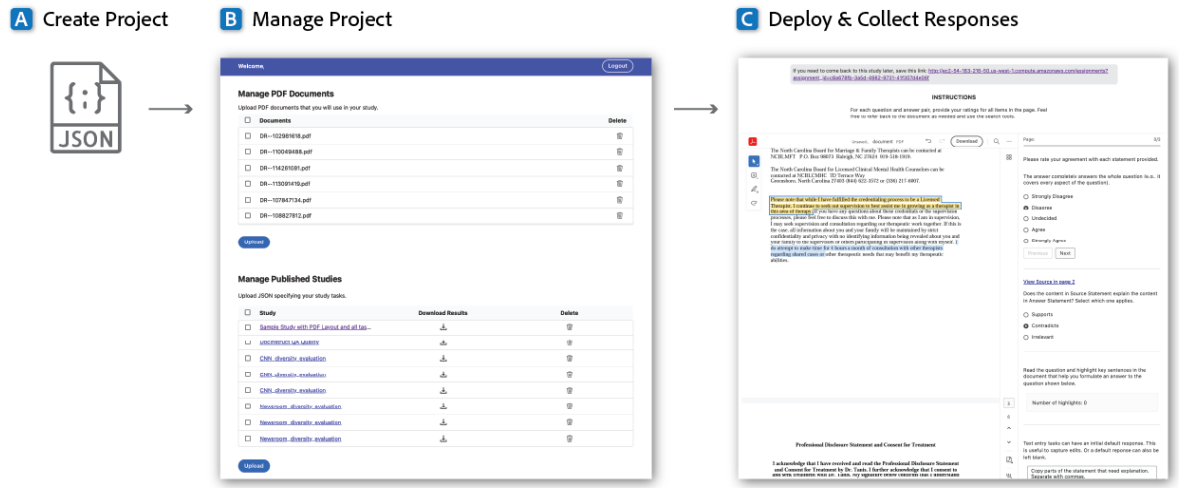


Figure 1: ATLAS supports researchers in programmatically creating a user interface for data collection that is ready to share with annotators. A) To start a project, the researcher first defines the data collection goals and tasks using the ATLAS toolkit which results in a JSON file. B) The researcher uploads the project JSON to the administrator interface which generates the data collection interface and a shareable URL. C) Project annotators access the URL to complete the specified data collection tasks.

ing and instrumenting PDF readers to capture rich interaction data while maintaining good usability.

Recently, several tools have been developed specifically to enable PDF annotations. PDFAnno (Shindo et al., 2018) and PAWLS (Neumann et al., 2021) are two representative tools, both of which enable annotating the PDF content, such as drawing and labeling bounding boxes for various PDF elements including texts, charts, headings, and captions. However, these existing tools only support *annotating the PDF content* and cannot support *collecting human* data based on the PDF such as readers' interactions with the PDFs including highlights, comments, and questions, nor can they support evaluating AI systems' outputs based on PDFs such as question answering quality and hallucination (Liu et al., 2023a; Li et al., 2023; Rawte et al., 2023). To enable the next-generation AI systems that collaborate with human users and readers on digital documents, we first need a tool that can aid the collection of large-scale and rich human data based on PDFs for training and evaluating document-based models.

### 1.1 Contributions

We propose ATLAS, a system for collecting PDF-centric human-interaction data. ATLAS complements existing data collection tools for PDF-based user interaction data collection. Unlike many existing tools that do not support collecting PDF-based data, ATLAS incorporates a native PDF viewer in its user interface, enabling the collection of various fine-grained interactions and annotations directly on the PDF file. Unlike existing PDF-based data collection tools that take a content-centric approach, focusing on annotations on and analysis of the PDF content, ATLAS takes a human-centric approach, focusing on user interaction data with the PDF that have become critical for user understanding, personalization, and model fine-tuning.

The ATLAS system consists of three main components: a visually consistent, native PDF-viewer integrated annotation interface that is dynamically and programmatically generated based on the specific annotation task, a toolkit, and an extensible schema to easily customize the data collection tasks for a variety of purposes, including annotation and interaction collection directly on PDFs; and a collection utilities for processing, analyzing, and visualizing the collected data, such as re-rendering the PDF annotations such as comments and highlights for further investigation.

The ATLAS's design makes it general, scalable, consistent, and easy to use. With ATLAS, researchers can create various data annotation tasks without writing a single line of UI code, enabling them to focus on the task design itself. And with ATLAS, task designers can then deliver the task they created at scale to hundreds of data annotators through a single, consistent UI that helps improve the annotators' experience and thus data quality (Tourangeau and Smith, 1996; Strong et al.,

1997; Bowling, 2005). A few examples, among the numerous data tasks that ATLAS can support, include PDF-based question-answering data collection, attribution data collection and evaluation, and reading behavior data collection, some of which have already been deployed in the wild to perform real-world PDF-based user interaction data collection and evaluations.

To summarize, we make the following contributions in this paper:

**[C1]** We propose ATLAS, the first-of-its-kind system for PDF-based human interaction collection, which existing data annotation tools cannot support.

**[C2]** We outline the architecture of ATLAS, which includes a programmatically generated user interface, a toolkit for creating a data collection task, and a suite of utilities for processing and analyzing the collected data. ATLAS's design makes it general, extensible, scalable, and easy to use. We also open-source ATLAS to support future research efforts.

**[C3]** We demonstrate via several concrete use cases to showcase ATLAS's wide applicability in real-world data annotation and evaluation scenarios.

## 2 Prior Work

Most existing data annotation tools (Stenetorp et al., 2012; Wei et al., 2013; Ogren, 2006; Yimam et al., 2014; Kummerfeld, 2019; Mayhew and Roth, 2018) support data collection for many data modalities and tasks except for PDFs. A few tools (Neumann et al., 2021; Lo et al., 2023) support annotating content in PDFs but do not support collecting human interaction data with the PDFs. ATLAS bridges the gap with a suite of features to support exactly these two scenarios, complementing the already vast landscape of existing annotation tools and software.

Because PDFs are designed to be read-only and immutable, providing support for native PDF annotations can be challenging. Therefore, a related line of efforts focuses on "morphing" the PDFs into other formats for easier annotation and processing. For example, Wang et al. (2021) proposed a method to parse PDFs into HTML format. However, converting PDFs to other formats usually loses some aspects of its original appeal, such as the persistence of its visual elements and layouts. It is well known the usability of the data annotation interface and presentation of data can impact user perception and thus data quality (Wobbrock et al., 2021; Spillane et al., 2018; Hausman and Siekpe, 2009; Coleman et al., 2008; Sonderegger and Sauer, 2010). Therefore, when collecting human interaction data with PDFs, using PDFs directly as part of the data collection and evaluation process is highly preferable. ATLAS enables this by providing a consistent UI for the annotator with a native, integrated PDF viewer capable of collecting fine-grained user interactions.

Many applications claim to perform intelligent tasks on PDFs such as search and retrieval, question answering, and summarization.[3] However, these applications are typically not transparent in evaluating how well they perform in these tasks and compare to competitors. Publicly available evaluation datasets and benchmarks are indispensable. Many existing datasets and benchmarks are text-only (Fabbri et al., 2021; Liu et al., 2023a; Kamalloo et al., 2023a), overlooking PDF documents. Authors of a few recent works collect PDFs as part of the dataset (Gu et al., 2024; Zhong et al., 2019; Li et al., 2020; Pfitzmann et al., 2022; Cheng et al., 2023), but the focus of these works are either annotations on the *content* of the PDFs and rarely on the *human interactions* with the PDFs (Lee et al., 2023). ATLAS provides an opportunity to enable easier and larger-scale collection of PDF-based human interaction data to benefit future developments of AI systems for documents and benchmark the progress. Some of the existing works are already empowered by ATLAS. For example, Saad-Falcon et al. (2023) developed a model for question answering over long, structured PDFs, in which ATLAS was central to collecting human data for evaluation.

## 3 ATLAS Design Choices

ATLAS supports researchers in programmatically creating a user interface for data collection that is ready to share with annotators. ATLAS scaffolds frontend creation and backend database management so that researchers can focus on the content needed for their data collection project. Figure 1 shows an overview of using ATLAS. The next sections describe each of ATLAS's user interface and design choices. A demo video is available in the

[3] Examples include https://www.chatpdf.com/, https://chatwithpdf.ai/, https://askyourpdf.com/, https://pdf.ai/, and https://chatdoc.com/

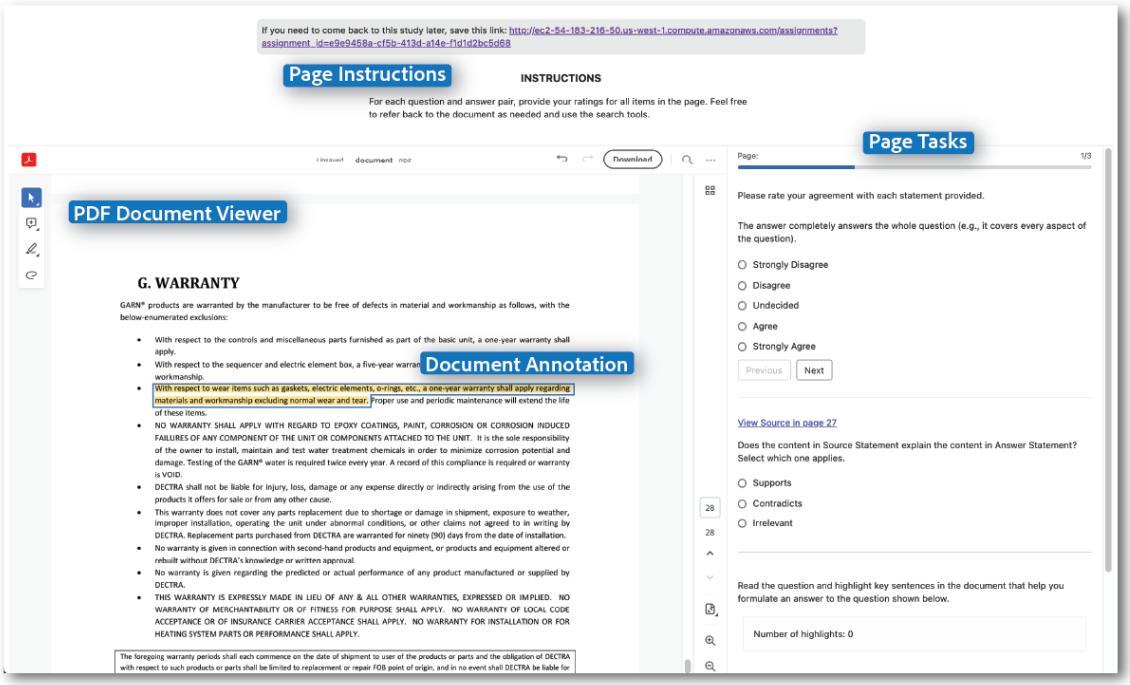


Figure 2: An example of the ATLAS annotator interface. Researchers can customize *Page Instructions* for each project. The *PDF Document Viewer* functions like a typical reader including common interactions such as scroll, zoom, text selection, and search. Researchers can include any number of tasks in the *Page Tasks* panel; these include survey tasks such as text entry, rank order, radio group, and checkboxes.

project GitHub repository[4] and in the supplementary material.

### 3.1 Toolkit Overview

ATLAS provides a toolkit library that allows researchers to define a data collection project programmatically. This programmatic approach provides flexibility in handling large amounts of data. A data collection project in ATLAS consists of the following components:

1. *Groups*: researchers can assign annotators to different annotation groups that contain different data collection tasks. For example, a researcher might have an evaluation where one group of annotators is only exposed to a control condition, and a second group is only exposed to a test condition.

2. *Pages*: an ATLAS project can contain any number of pages that are presented to annotators and guide the flow of the data collection project. Pages contain a PDF reader and relevant tasks for annotators to complete.

3. *Tasks*: a page can contain any number of data collection tasks (Figure 2). These can be either required or optional for annotators to proceed through the data collection. Section 3.2 details all the tasks supported.

### 3.2 Creating Instructions and Tasks

Once a researcher has defined a project goal, the first step in using ATLAS, is defining the instructions and tasks for data collection using the ATLAS toolkit, which results in a JSON file (Figure 1A).

Researchers can include project-level, page-level, and task-level instructions. These allow researchers to provide proper context and guidance to annotators providing responses. For example, start instructions can include annotation examples for annotators to review before starting any data collection tasks. Task-level instructions can also include a document source that scrolls the PDF to a specific location in the PDF. Document sources can be useful if a task requires the user to pay attention to a specific section or statement in the document.

For data collection, the ATLAS toolkit currently supports the following data:

1. Bounding boxes of PDF annotations (i.e., page highlights, free-form drawings, com-

[4] https://github.com/frictionlessweb/documentstudies/

ments) and underlying content (i.e., text, images, tables)

2. Timestamps of reading behavior analytics (e.g., document scroll position, zoom level, clicks, text search). Table Appendix B describes all behavior analytics that are currently supported.

3. Survey user responses (i.e., text entry, rank order, radio group, and checkbox group)

### 3.3 Managing and Deploying a Project

After an ATLAS project is created, the next step is uploading the project to the ATLAS administrator interface (Figure 1B). Additionally, a researcher uploads any PDF documents that are used in the data collection project. Once uploaded, ATLAS automatically generates a URL that is ready to be shared with annotators to begin the data collection tasks. Each project has a unique URL and a researcher can create any number of projects. All projects are listed in the administrator panel. Once the data collection is complete, the researcher can download the results for each project. Results are exported in a JSON file.

### 3.4 Collecting and Analyzing Responses

Researchers share a URL with annotators where they can begin working on the data collection tasks. Once an annotator begins a project, a unique URL is generated for that annotator. This allows an annotator to take breaks between tasks and/or return to the project at different times without losing their progress on completed responses. A demo URL is accessible through the project's GitHub repository.

A researcher can download responses for a project at any time. Downloads include any in-progress/partial responses as well as completed responses. The ATLAS toolkit provides functions to help researchers aggregate responses from users. Additionally, post-processing functions enable researchers to map PDF annotation bounding boxes to semantic content (i.e., text, images, tables).

## 4 Implementation

*Backend:* ATLAS is implemented using the web-app framework, Ruby on Rails[5]. PostgreSQL is used for the database. There are three key data models: 1) A *Document Table* handles file metadata and nomenclature. 2) A *Project Table* encapsulates comprehensive details about each project. 3) A *Project Assignments Table* tracks completed annotator responses.

*Toolkit Processing Libraries:* To specify a data collection project, researchers create a JSON file specifying the instructions and tasks. The ATLAS toolkit provides a set of Python functions that allow customizing the tasks. All instructions can be formatted using Markdown syntax. Once complete, the JSON file is uploaded to the application along with the required PDF documents. The toolkit also provides functions to aggregate results into readable data tables and post-process PDF annotations.

*Frontend:* The user interface is implemented using React with state management across the application governed by the context API. There are two main interfaces, the researcher interface, and the annotation interface. PDFs are rendered using Adobe's PDF Embed API.[6]

## 5 Example Use Cases

To showcase our system's broad applicability and impact, we review three example projects that have leveraged ATLAS's capabilities to meet research goals.

**Document-Grounded Question Answering (QA)** This project aims to collect pairs of questions and answers based on documents as well as attributions (texts, tables, images, or charts in the document) that help explain the answer. This dataset is similar to the Natural Questions dataset on Wikipedia articles (Kwiatkowski et al., 2019) but on a wider set of PDF documents from different areas of domain expertise. This dataset can be a crucial first step in fine-tuning models for automatic question answering, question generation, and source attribution grounded on the source document. The dataset is also critical for evaluating and benchmarking the performance of these document-based models. ATLAS enables researchers to create a custom study using openly licensed documents and annotation tasks where annotators author (potentially multiple) question-and-answer pairs for each document. Annotators can also highlight the source texts or other content in the PDF as attribution. ATLAS helped standardize, streamline, and significantly scale the

[5] https://rubyonrails.org

[6] https://developer.adobe.com/document-services/apis/pdf-embed/

data collection effort to over 3,000 unique documents and more than 50 data annotators, resulting in a dataset of over 20,000 question-and-answer pairs along with their attributions and forming a solid foundation for future model fine-tuning and benchmarking.

**Document-Grounded QA Evaluation** This project aims to scientifically and systematically evaluate large language models' (LLMs) document-based QA capabilities using metrics including answer quality, attribution, and bias. In particular, accurate and precise attribution is an essential feature to improve LLMs' trustworthiness when answering questions based on source documents (Liu et al., 2023a). Existing tools and literature focus on evaluating attributions based primarily on free-form texts but rarely on evaluating PDF-based QA attributions (Kamalloo et al., 2023b; Huang et al., 2023; Yue et al., 2023). However, providing a faithful interface that represents how real readers read digital documents for evaluation PDF-based QA attributions is important for ensuring evaluation consistency and quality (Kwon et al., 2014). ATLAS supports such evaluation by providing a consistent annotation UI with a natively integrated PDF viewer where attributions are shown as highlights directly in the PDF. Researchers can flexibly and programmatically create the evaluation task to include different types of quality metrics that they want to collect without the need to alter the UI. ATLAS has supported several rounds of attribution evaluation using two approaches. In one approach, annotators directly rate the quality of machine-generated attributions in terms of precision and recall. In a second approach, annotators provide their own attributions, and agreement between human- and machine-generated attributions is computed. In addition to attribution, ATLAS has also been used for evaluating model answering quality and bias. As an example, Saad-Falcon et al. (2023) have employed ATLAS to evaluate their novel automated QA methodology for long, structured documents.

**Digital Reading Behavioral Data Collection** Behavioral data on how people read and interact with digital documents can help deepen the understanding of reading patterns, improve the design of reading applications, and develop better personalization technologies for a more delightful digital reading experience (Rajendran et al., 2018; Wallace et al., 2022; Maity et al., 2017). Such data is typically proprietary and there exists no open-source tools to support the collection of such data. ATLAS aims to change the landscape by providing the capability to collect fine-grained implicit reading behaviors such as temporal mouse-over patterns, clicks, scrolls (direction and speed), search queries, comments, and highlights. ATLAS enables a non-intrusive way to collect such data in a reading interface that closely represents common software for consuming PDFs such as Adobe's Acrobat Reader and Apple's Preview, improving the representativeness of such behavioral data collected using ATLAS to real-world reading patterns. An ongoing study leverages ATLAS to collect one-of-a-kind, open-source, large-scale reading behaviors from professionals across various industries in the hope of unlocking future research in studying, analyzing, and improving digital reading experiences.

## 6 Conclusions

In this paper, we presented ATLAS, an open-source system for collecting PDF-centric human interaction data. ATLAS complements existing data collection tools by focusing on PDF-based user interactions and supports a wide range of interaction data collection tasks, such as question-and-answer pairs, QA attributions, and reading annotations and behaviors. It features a programmatically generated user interface, a toolkit for creating data collection tasks, and a suite of utilities for processing and analyzing the collected data. We demonstrated ATLAS's capabilities and applicability through several real-world use cases. We believe that ATLAS will be a valuable tool for researchers and practitioners working with PDF-based human interaction data, and we hope that it will enable new and exciting research in this area.

## Broader Impacts and Limitations

ATLAS holds the potential to significantly impact document-based AI advancements, user experience design, and research accessibility. By enabling the collection of rich human interaction data on PDFs, it paves the way for more sophisticated AI models that understand and interact with documents, leading to improved question-answering, summarization, and personalization. This democratization of data collection empowers researchers and practitioners alike, fostering new avenues for document-based technology development. Furthermore, the data collected through ATLAS can shed light on

user reading patterns, informing the design of more intuitive reading interfaces and navigation tools. Additionally, its unique capability for capturing PDF interactions allows for rigorous benchmarking and evaluation of document-based AI, fostering transparency and trust in these models.

However, it's important to acknowledge ATLAS's limitations. First, although ATLAS can be extended to data collection tasks beyond PDFs, it is currently limited to the formats of documents it can support. Expanding to include additional document formats, like Word documents or ePub files, would broaden its utility. Second, collecting user data carries ethical considerations. Robust security measures and data anonymization are essential to ensure participant privacy and trust. Third, scalability and efficiency remain to be tested and ensured, as handling extra large datasets and complex tasks can strain system resources. Optimizing the platform for smoother performance such as high-performing databases and load balancing will be crucial for supporting even large-scale research projects. Finally, any data collection effort might inadvertently introduce bias. Researchers must be mindful of these potential biases and employ appropriate mitigation strategies to ensure the collected data accurately reflects real-world interactions.

By addressing these limitations and continuously evolving, ATLAS strives to be a valuable tool for responsible and ethical data collection, ultimately fostering the development of trustworthy and impactful document-based AI technologies that benefit all users.

## References


Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Ann Bowling. 2005. Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, 27(3):281–291.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. 2023. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15138–15147.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Renita Coleman, Paul Lieber, Andrew L Mendelson, and David D Kurpius. 2008. Public life and the internet: if you build a better website, will citizens become engaged? *New media & society*, 10(2):179–201.

Andrea Cuadra, Hansol Lee, Jason Cho, and Wendy Ju. 2021. Look at me when i talk to you: A video dataset to enable voice assistants to recognize errors. *arXiv preprint arXiv:2104.07153*.

Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Jiuxiang Gu, Xiangxi Shi, Jason Kuen, Lu Qi, Ruiyi Zhang, Anqi Liu, Ani Nenkova, and Tong Sun. 2024. ADoPD: A large-scale document page decomposition dataset. In *The Twelfth International Conference on Learning Representations*.

Angela V Hausman and Jeffrey Sam Siekpe. 2009. The effect of web interface features on consumer online purchase intentions. *Journal of business research*, 62(1):5–13.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023a. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023b. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution.

Jonathan K. Kummerfeld. 2019. SLATE: A super-lightweight annotation tool for experts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12, Florence, Italy. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Ohbyung Kwon, Namyeon Lee, and Bongsik Shin. 2014. Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, 34(3):387–394.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-In Lee, and Moontae Lee. 2023. QASA: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19036–19052. PMLR.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment.

Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. 2023. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore. Association for Computational Linguistics.

Suman Kalyan Maity, Abhishek Panigrahi, and Animesh Mukherjee. 2017. Book reading behavior on goodreads can predict the amazon best sellers. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17. ACM.

Stephen Mayhew and Dan Roth. 2018. TALEN: Tool for annotation of low-resource ENtities. In *Proceedings of ACL 2018, System Demonstrations*, pages 80–86, Melbourne, Australia. Association for Computational Linguistics.

Mark Neumann, Zejiang Shen, and Sam Skjonsberg. 2021. PAWLS: PDF annotation with labels and structure. 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 258–264, Online. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Philip V. Ogren. 2006. Knowtator: A protégé plug-in for annotated corpus construction. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*, pages 273–275, New York City, USA. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John

Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3743–3751, New York, NY, USA. Association for Computing Machinery.

Ramkumar Rajendran, Anurag Kumar, Kelly E Carter, Daniel T Levin, and Gautam Biswas. 2018. Predicting learning by analyzing eye-gaze data of reading behavior. *International Educational Data Mining Society*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models.

Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, David Seunghyun Yoon, Ryan A. Rossi, and Franck Dernoncourt. 2023. Pdftriage: Question answering over long, structured documents.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Hiroyuki Shindo, Yohei Munesada, and Yuji Matsumoto. 2018. PDFAnno: a web-based linguistic annotation tool for PDF documents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press, London, England.

Andreas Sonderegger and Juergen Sauer. 2010. The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied ergonomics*, 41(3):403–410.

Brendan Spillane, Séamus Lawless, and Vincent Wade. 2018. Increasing and decreasing perceived bias by distorting the quality of news website design. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*, pages 1–13.

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 at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Ashley Still. 2020. Adobe unveils ambitious multi-year vision for pdf: Introduces liquid mode.

Diane M. Strong, Yang W. Lee, and Richard Y. Wang. 1997. Data quality in context. *Communications of the ACM*, 40(5):103–110.

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.

Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-first dataset creation: Lessons from constructing a dataset for teachable object recognition with blind and low vision data collectors. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–12.

Roger Tourangeau and Tom W. Smith. 1996. Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, 60(2):275.

Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B Miller, Jeff Huang, and Ben D Sawyer. 2022. Towards individuated reading experiences: Different fonts increase reading speed for different individuals. *ACM Trans. Comput. Hum. Interact.*, 29(4):1–56.

Lucy Lu Wang, Isabel Cachola, Jonathan Bragg, Evie Yu-Yen Cheng, Chelsea Haupt, Matt Latzke, Bailey Kuehl, Madeleine van Zuylen, Linda Wagner, and Daniel S. Weld. 2021. Improving the accessibility of scientific documents: Current state, user needs, and a system solution to enhance scientific pdf accessibility for blind and low vision users.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 41(W1):W518–W522.

Jacob O Wobbrock, Lara Hattatoglu, Anya K Hsu, Marijn A Burger, and Michael J Magee. 2021. The goldilocks zone: young adults' credibility perceptions of online news articles based on visual appearance. *New Review of Hypermedia and Multimedia*, 27(1-2):51–96.

Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. 2014. Automatic annotation suggestions and custom annotation layers

in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Baltimore, Maryland. Association for Computational Linguistics.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. (inthe)wildchat: 570k chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis.

## A Toolkit JSON Schema

The output of the ATLAS toolkit is a JSON file which is then uploaded to the ATLAS interface. The JSON file defines the data collection instructions and tasks. Below is a snippet of the schema for specifying a project:

```
{ "schema_version": "v0",
  "metadata": {
    "name": str,
    #... any other needed fields can be added ...
  },
  "start_instructions": str,
  "end_instructions": str,
  "groups": [ 0 ],
  "group": null,
  "page_index": 0,
  "content": {
    "0": { "pages": [
      { "id": str,
        "page_layout": str <pdf_layout, text_layout>,
        "instructions": str,
        "document_id": str,
        "hide_previous_button": bool <default false>
        "save_pdf_interactions": bool <default false>
        "tasks": [ #... list of tasks ... ] },
        #... any number of tasks can be added ...
    ] }
  }
}
```

## B Reading Behavior Analytics

ATLAS supports data collection of reading behavior analytics. These events are captured as timestamps and include the following:

1. Current active page: Changes to the page in view
2. Text copy: On copying text from the document
3. Text search: When the user searches for any text via the document search field
4. Zoom level: When zoom-in/out actions are performed from the page control toolbar
5. Page click: When a user clicks on any document page
6. Page double click: When a user double clicks on any document page
7. Mouse enter/leave: The mouse pointer enters/leaves any page
8. Annotation added: A new annotation is added to the document
9. Annotation clicked: An existing annotation is clicked
10. Annotation updated: An existing annotation is updated
11. Annotation deleted: An annotation is deleted
12. Annotation mouse over or mouse out: The mouse pointer moves over/out of any annotation
13. Annotation selected or unselected: Any existing annotation is selected/unselected
14. Annotation count: Total number of document annotations updated whenever a new annotation is added or any existing annotation is deleted