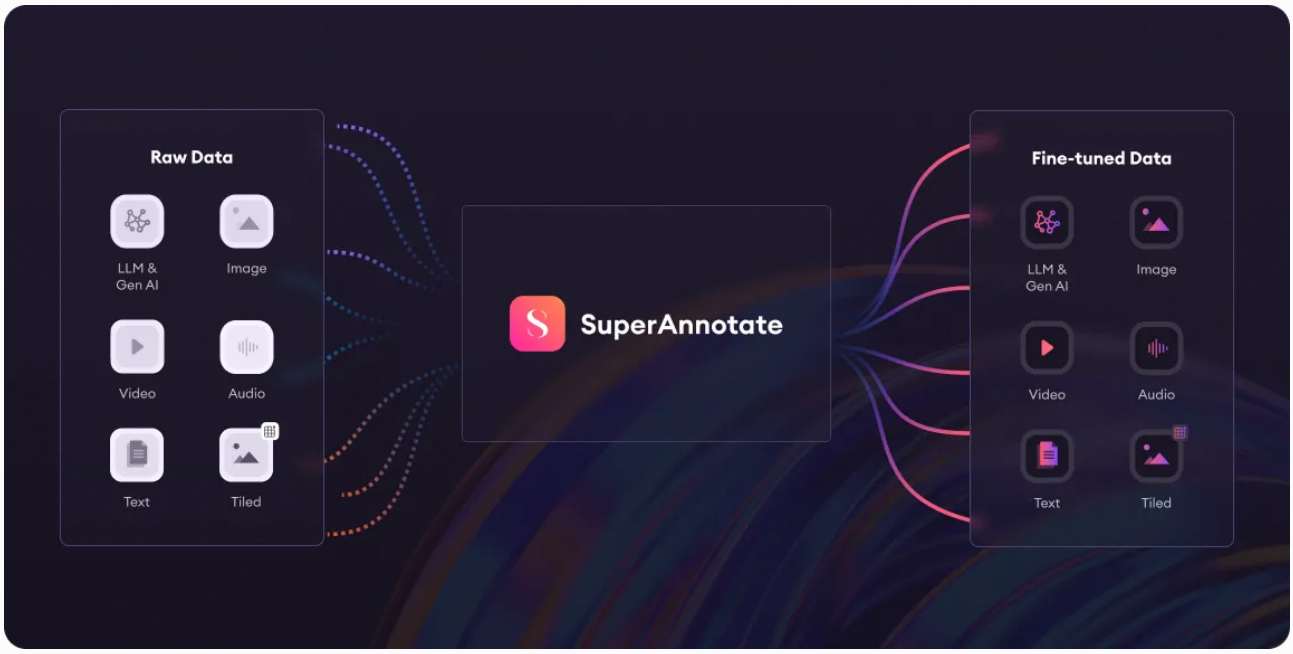
SuperAnnotate

The CTOs guide to SFT and RLHF projects

How to successfully execute SFT and RLHF data collection and LLM evaluation projects at scale



Introduction

Since the advent of ChatGPT in late 2022, the urgency to develop and deploy effective LLMs has intensified, with enterprises and startups eager to leverage their capabilities. A critical challenge in this task is aligning LLMs with human expectations, essentially tailoring these models to behave in ways users find most beneficial and intuitive.

Addressing this challenge involves sophisticated methods like Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), which depend heavily on large, incredibly high-quality, human-generated datasets. Due to this reliance on large-scale, highly-skilled workforces, creating such datasets poses significant hurdles, encompassing not just the volume of data but its quality, diversity, and ethical implications.

This white-paper draws upon SuperAnnotate's extensive experience in assisting major companies with LLM dataset development, offering insights to help teams that are building or fine-tuning models with navigating the complexities of dataset creation for LLM training. We delve into critical aspects such as achieving alignment across data creation teams, handling sensitive or controversial content, the pitfalls of crowdsourcing data, strategies for scaling quality, the benefits of ongoing model evaluation, and why you should intentionally break your model.

Copyright © 2024 SuperAnnotate AI, Inc. All rights reserved.

Table of Contents

Style Guide

Anyone interacting with a large language model has probably realized that different models behave differently. Some are more chatty, and others are more assistant-like. Some will have no problem answering controversial questions, whereas others might take a more conservative approach or decline to answer entirely. Different styles are appropriate for different use cases; what works well for a language model intended for character role-play might differ significantly from what is suitable in a model created for an enterprise environment.

As a model creator, you may have well-developed ideas about how you would like the model to behave. Yet, you will most likely be disappointed if you turn to a crowdsourcing platform, describe this in a brief document, and then train the model on the resulting data. Every person involved in writing data (known as a trainer) will interpret your instructions differently; some responses will be of higher quality than others.

Standard American English

Definition

The language should reflect Standard American English at an 8th-grade level, ensuring that the Response is accessible to a broad audience. It should use contractions such as can't and don't instead of cannot and do not ...

Formatting

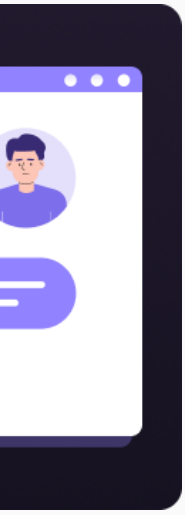
Definition

The Response should be formatted in a way that aids readability and understanding. Lengthy paragraphs and lists should be separated by line breaks. Whenever order matters, lists should be numbered. Otherwise, lists should ...

Getting on the same page

When engaging with clients, one of the first steps we take is to create a comprehensive style guide. This style guide outlines the intended model style and behavior in excruciating detail, from how the model refers to itself to what type of paragraph breaks and bullet lists to use to detailing the outline and style of responses. Our style guides can often reach 30-40 pages in length and continue to grow throughout an engagement.

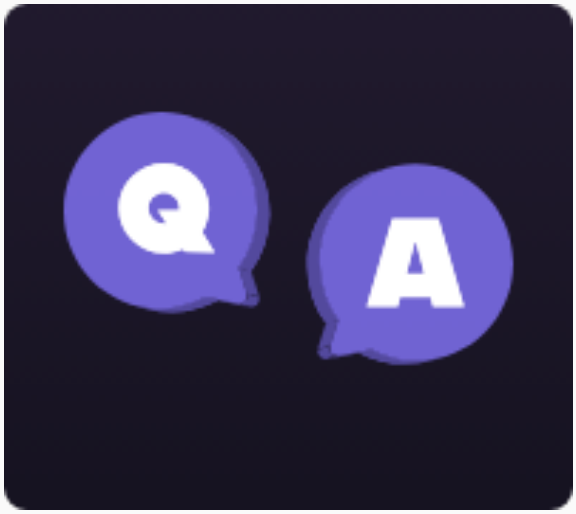
A comprehensive style guide is crucial for directing the training process. It is intended to ensure that everyone involved in the project interprets the intended style and behavior similarly. Creating a style guide is a collaborative effort between the data collection project experts and the large language model team. It often requires input from linguists, domain experts, and sometimes lawyers to ensure it covers every possible nuance.



Clarity

Definition

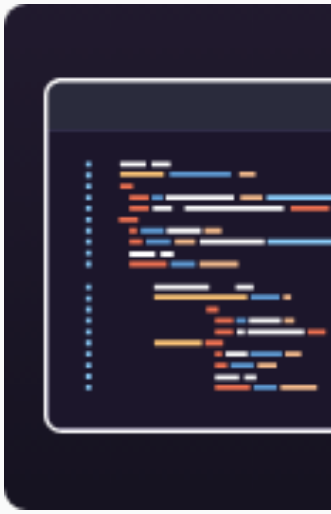
The Response should provide information that is easy to understand, concise, and free of ambiguity. It should be straightforward and focus on delivering the message in the simplest ...



Tone of Voice

Definition

The Response should adhere to a professional and polite standard of communication without being overly long, short, technical, pedantic, or stiff (unless specified in the User Request). The language should be ...



Workforce Management

As large-scale data collection and model evaluation projects transition from the initial phases to expansive operations engaging a vast workforce tasked with generating tens to hundreds of thousands of data items, ensuring the maintenance of quality and consistency becomes a formidable challenge. Creating a detailed style guide lays a foundation for being successful. However, ensuring adherence to this guide and factual accuracy across a growing workforce generates its own challenges.

Radiologist

Adam Scott

Linguist

John Williams

Medical doctor

Edward Johnson

PhD student

Lucas Roberts

Lawyer

Harper Thompson

Developer

Sophia Miller

Financial specialist

Olivia Jones

Finding Experts

When building datasets for SFT and RLHF for LLMs, outsourcing is often a necessity due to the extensive scope of work involved. Identifying a reliable workforce provider is crucial, as the magnitude of LLM projects means that making a wrong choice could lead to significant financial losses and project delays. To mitigate these risks and foster a successful collaboration, it is essential to consider the following four key factors during the selection process.

1

Domain Expertise
LLM data creation is markedly different from other dataset tasks. Teams without direct LLM experience may overlook crucial intricacies that impact model performance. For projects involving specialized topics, possessing relevant domain expertise is equally critical.

2

Project and QA Management
Combining project and quality assurance (QA) management with the team is key to implementing advanced quality control and training processes necessary for success. For an illustration of this concept, refer to the section titled "The Risk with Crowdsourcing."

3

Collaboration
A collaborative partner who can serve as an extension of your team enables you to leverage their expertise to develop and enhance the style guide continuously over time. Initiating a short pilot project is an effective strategy to determine the effectiveness of the collaboration.

4

Scalability
Quickly increasing output without affecting quality is a significant challenge. Be cautious with companies that promise very high scalability. Such rapid expansion often leaves insufficient time for training and quality assurance (QA), potentially resulting in subpar quality.

The risk with crowdsourcing

Crowdsourcing presents a tempting solution for rapidly expanding Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) data collection efforts. However, its capacity to scale quickly is balanced by significant challenges, especially in managing quality and protecting intellectual property. While a comprehensive style guide is important for standardizing instructions, its distribution among a vast network of crowd workers introduces the risk of exposing sensitive information in a fiercely competitive arena.

> **“Often, crowd workers became unreachable after submitting their initial work, complicating any attempts at iteration.**

Experience has taught us that relying solely on crowdsourced labor and freelancer platforms for data collection can lead to inconsistent quality. Initially, our projects adopted the approach of hiring freelancers and crowd workers, sending the instructions, and awaiting the returned dataset. While this method yielded high-quality data in some instances, it frequently fell short. Moreover, we encountered difficulties getting workers to refine their contributions based on our feedback.

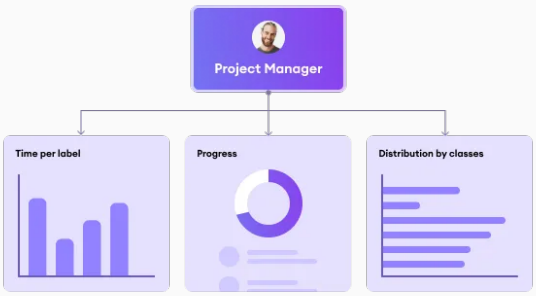
Often, crowd workers became uncontactable after submitting their initial work, complicating any attempts at iteration. Additionally, the limited communication tools provided by many crowdsourcing platforms hampered our efforts to provide clear, timely feedback.

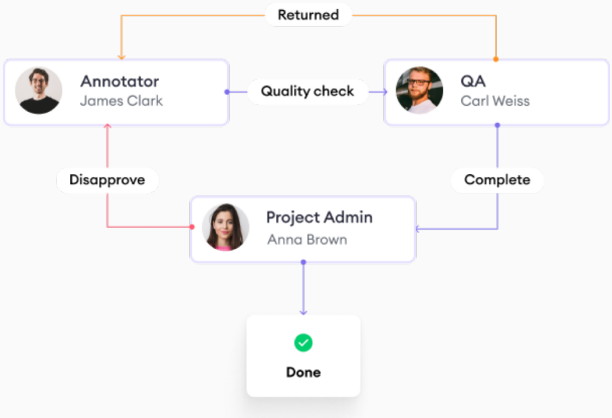
We now prefer to employ data trainers directly to address these issues, enabling us to create a more controlled and efficient environment. By taking this approach, we can more effectively manage the workforce, enforce non-disclosure agreements, and implement sophisticated project setups alongside rigorous QA processes. This direct-hiring model is critical in following the structured QA and annotator development strategies previously outlined, and helps ensure data collection efforts scale without compromising the integrity and quality of the dataset.

The three Cs of Data Quality

Clarity

Adopting an in-house or fully managed trainer model enables more in-depth project analytics. It allows for nuanced performance tracking, training needs identification, and monitoring of quality assurance trends. This clarity around project status elevates the overall quality of the data collected.



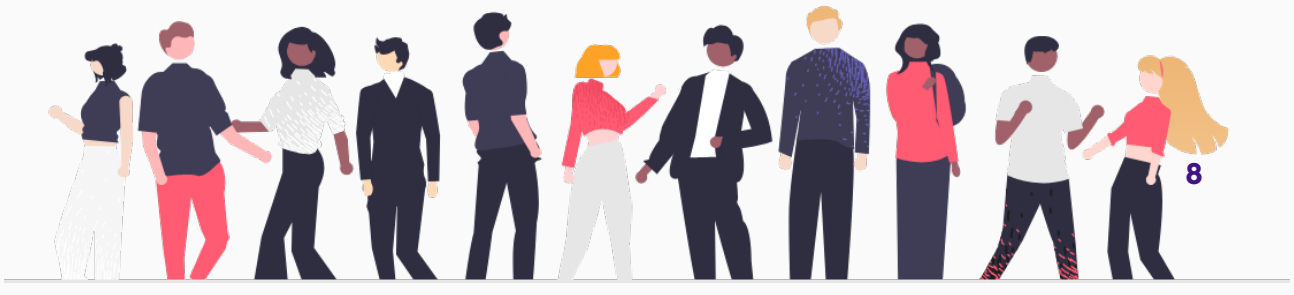


Communication

Moving away from the impersonal and often restrictive communication channels of crowdsourcing platforms has allowed us to have a more direct dialogue within our teams. This approach enables us to quickly address questions, provide feedback, and adjust requirements, ensuring every team member is aligned with the project's goals.

Commitment

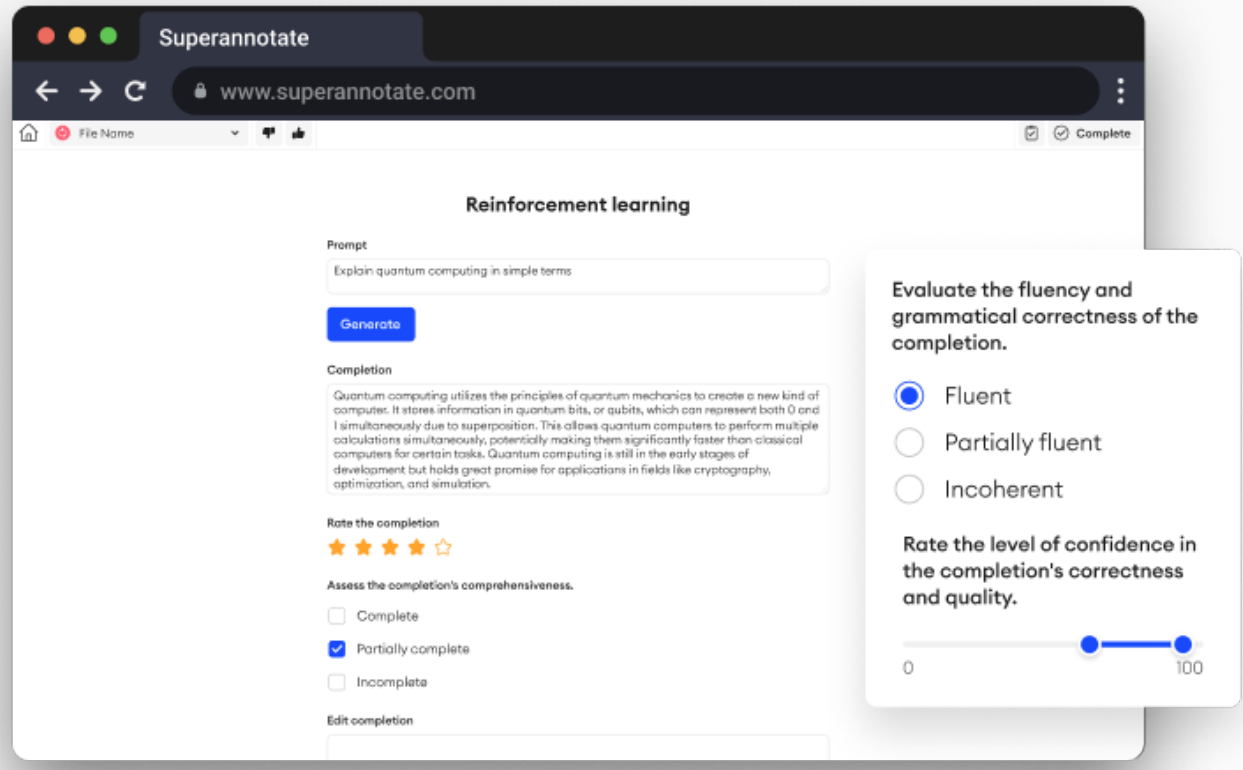
Non-crowd-sourced trainers are more committed and driven by a clear understanding of their role in the larger objective and the opportunity for professional growth within the organization. A dedicated team is more likely to go above and beyond, ensuring that the data meets the required standards and exceeds them wherever possible.



Evaluating Models

In its most basic form, the evaluation of large language models can consist of having users write a prompt that gets sent to two different models. The user then ranks which of the resulting responses is the best. This approach facilitates the calculation of comparative metrics, such as win rates or ELO scores, which provide a quantitative basis for assessing model performance. While beneficial for a high-level comparison, pairwise evaluation lacks the granularity to guide where to best focus efforts to improve the model.

> **“Pairwise evaluation lacks the granularity to guide where to best focus efforts to improve**



Criteria

Establish a set of evaluation criteria. The criteria might include measures of toxicity, the percentage of hallucinated content, adherence to multiple aspects of the style guide, etc.

Prompts

Creating a dataset specifically for evaluations helps prevent overestimating model performance by using prompts not seen during training. If evaluation prompts are used to make training decisions, consider creating a validation and test dataset.

Regularity

Conduct blind evaluations regularly to allow developers to identify when specific performance benchmarks have been achieved, pause data collection in certain areas, and reallocate resources toward enhancing other aspects of the model.

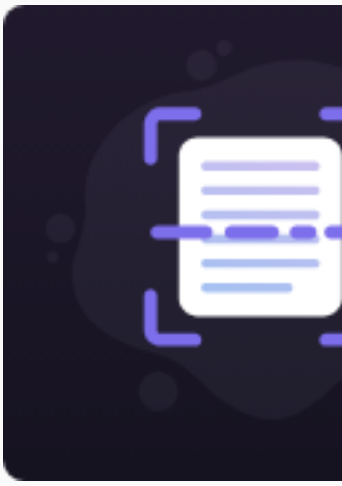
Tracking

We have seen instances where SFT managed to increase the performance on reasoning benchmarks but also increased model toxicity. Tracking various performance indicators can uncover and address these types of situations.

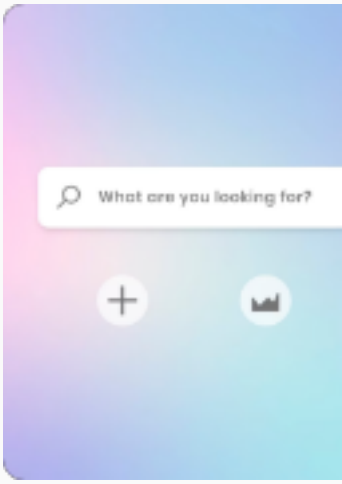
Will it break? How to think about red-teaming

Traditional evaluation is excellent for understanding how well a large language model performs under the intended uses, but it is not always enough to catch unintended behaviors. Several different model behaviors can be unintentional, the most prominent of which relates to how the model handles controversial or unsafe topics. Some models mistook safe queries for unsafe ones, refusing to answer questions about how to kill a car engine. Most recently, Google's Gemini model said that the issue of who was the worst of a well-known American businessman and a German dictator was nuanced and complex, and mechanisms intended to reduce bias resulted in the generation of inaccurate historical images.

Red teaming attempts to address and uncover a vast array of issues, and it can, as illustrated by the examples above, make it easy for unwanted aspects to slip through the cracks and into the released product. It is, therefore, essential to take care of this when designing and executing the red-teaming process.



In our experience there are five key factors that make or break red-teaming efforts



1 Allocating Sufficient Time

Invest at least two months and sufficient human resources to develop a comprehensive and varied set of prompts for each Red Teaming category, ensuring a thorough evaluation of the LLM's capabilities and weaknesses.

2 Detailed Evaluation Rubric

Construct a comprehensive rubric for assessing responses systematically. Areas of focus should include the occurrence of hallucinations, adherence to refusal policies, and other critical metrics.

3 Assembling a Skilled Team

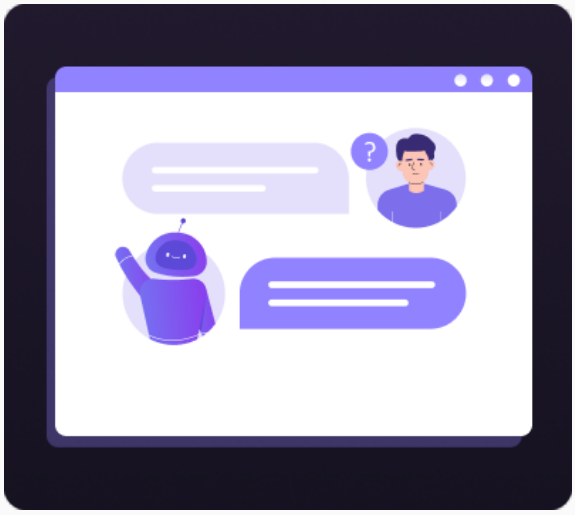
Hire a dedicated team of experienced AI writers, coders, mathematicians, etc, with specialized skills capable of crafting complex prompts and analyzing nuanced responses to challenge the LLM effectively.

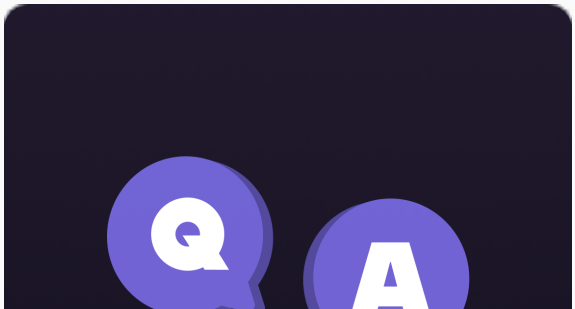
4 Statistical Analysis of Red Teaming Data

Conduct comprehensive reviews of all data collected from Red Teaming tests to identify patterns and trends. This can provide additional insights into the model behavior and potential biases.

5 Ensure Transparency and Documentation:

Maintain detailed records of Red Teaming methodologies, tests, and findings.

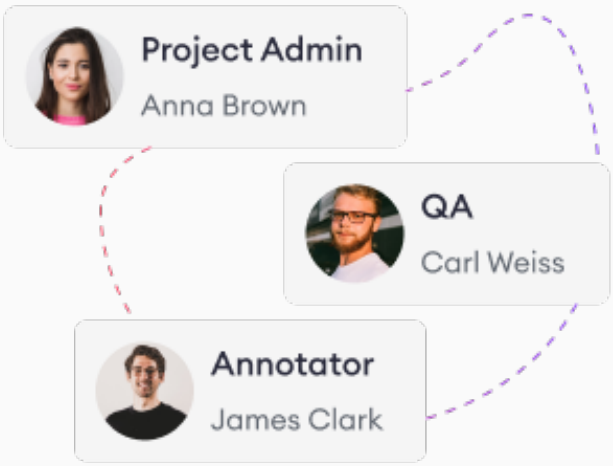




Why software is important

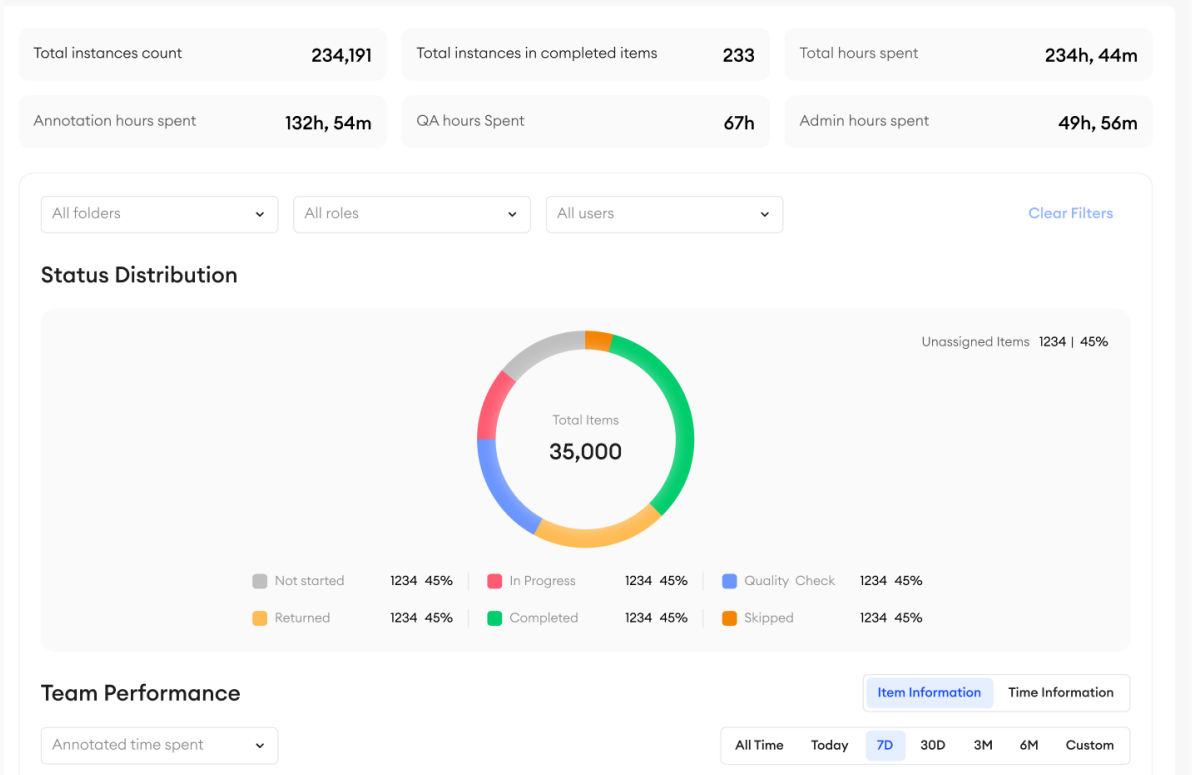
Integrated Workflow and Quality Assurance Tools

LLM dataset creation projects require extensive QA, routing of items between several people, and efficient communication. Without software that offers comprehensive workflow management functionalities and QA communication and enables efficient routing of tasks among annotators, reviewers, and experts, this quickly becomes unmanageable.



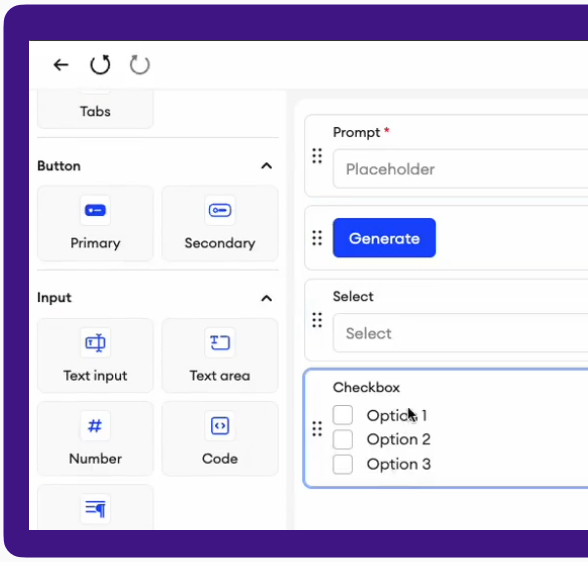
Comprehensive Visibility and Reporting

Without detailed visibility into project workflows, including the ability to query actions, track progress, and generate reports, it is impossible to monitor work distribution, assess productivity, and identify areas needing attention or improvement.



Customizable Interfaces

LLM Dataset Creation compromises tasks such as SFT, RLHF, evaluations, and red-teaming exercises. Furthermore, you may introduce new tasks or change requirements as new research is released. Software that allows for the customization of interfaces and support for multimodal data, including text, images, videos, and web content, removes the need to have system-developing staff constantly work on annotation tooling to keep up with project changes.

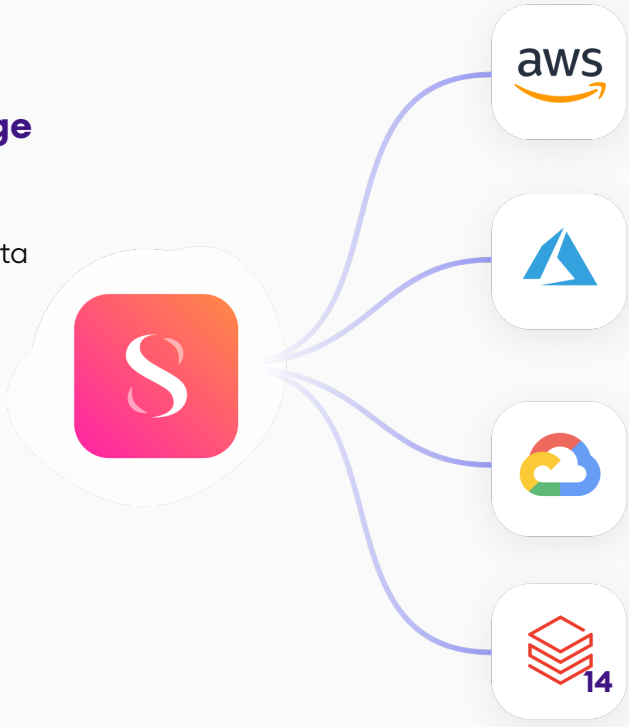


Ability to connect APIs

Integrating models directly in the training or evaluation pipeline can improve efficiency. However, this integration is only possible with software that supports functionalities like prompt submission, response collection, and performance analysis. Additionally, enhancing the feedback process through improved data quality and annotator guidance is difficult without integrating with external tools like grammar checkers or AI evaluation systems.

Integration with Data Storage Solutions

Ability to integrate efficiently with data storage solutions streamlines data management, ensuring seamless access to and from centralized data repositories, facilitating more accessible updates, and enhancing overall project scalability



Having a partner

Developing and fine-tuning a large language model is challenging, and the complexities involved in high-quality dataset creation and model evaluation are hard to understate. From crafting detailed style guides to managing extensive workflows for Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), each step requires meticulous attention to detail, deep expertise, and significant time and resource investment.

> **The complexities involved in high-quality dataset creation and model evaluation are hard to understate.**

1 Preparing the data collection

Building and defining processes and workflows from scratch is a monumental task, even with the guidance of this report. Style guide creation, defining evaluation criteria, and building an approach to red

2 Building and managing software

Developing in-house software is both time and resource-intensive. It requires a dedicated team of developers, constant updates to keep pace with evolving project requirements, and significant

3 Managing a workforce and quality

Opting to manage the workforce in-house necessitates a substantial investment in staffing, training, and ongoing development. This approach strains budgets and diverts focus from the primary objective of

4 Being the in-between

If you acquire a platform and hire a workforce from separate providers, your team can be left to navigate software issues, integration headaches, and communication barriers.

Choosing a partner with experience, a purpose-built platform, and dedicated management for an LLM-specific workforce offers a streamlined solution to several of the challenging aspects of LLM data collection projects.

By consolidating data creation and evaluation expertise, resources, and management under a single umbrella, teams can focus on their core objective of building LLMs while minimizing the operational burdens associated with complex project execution. This partnership not only accelerates the development timeline but also enhances the quality and effectiveness of the final model, ensuring that the LLM meets the high expectations of its intended users.

Experience

Leveraging the partner's expertise in navigating the nuances of LLM dataset creation and evaluations can save considerable time and help you avoid common pitfalls.

Purpose-Built Platform

Accessing a platform specifically designed for the purpose and equipped with the tools for data management, workflow automation, and quality control removes the need to build and maintain your own platform and provides tools to make the work more efficient, which might not have been feasible to build for an internal platform.

Managed Workforce

Benefiting from a dedicated, skilled workforce managed by the partner, eliminating the need for extensive in-house recruitment, training, and quality management efforts.

SuperAnnotate

Copyright © 2024 SuperAnnotate AI, Inc. All rights reserved.