Overview

Timeline: 2 Months

Members: Emily Gao, Andy Xu, Sadhvi Narayanan, AJ Matheson-Lieber, Ellie Lian, Felix Peng

Figma Prototype [here]

Reviewing Synthetically Generated Data: Inconsistent, Tedious, and Manual

Instructlab is an open source IBM project, training and fine tuning enterprise-level LLMs with synthetically generated data on their flagship watsonX AI product. By leveraging InstructLab, WatsonX allows enterprises to create custom models trained on highly specific datasets and can deliver specialized models that perform on tasks such as data analysis, customer service, etc.

Problem Space

The current integration does not include a consistent way to review the hundreds of sets of data that are being fed into the model, relying on the irregular manual review processes of each individual team. The Graphite Digital team was approached with the problem, "How might we allow teams of reviewers to efficiently and collaboratively approve or deny sets of synthetic data?"

Current Solutions and Pain Points

"If I'm unsure about the process or answer, I just hope that someone gets around to reviewing the question."
– IBM Developer

The current reviewal process includes retrieving synthetic data and transferring it into a CSV or JSON file where it will be reviewed — any indication of who's tackling a specific question-and-answer data set, or if it's already been reviewed, is arbitrarily determined by each team. The alternative — directly reviewing in the command line interface — was favored by only technical users, becoming a large barrier for non-technical reviewers. During our preliminary user research with IBM developers and reviewers, it quickly became clear that the lack of such a reviewing tool has caused setbacks in model training time and gaps in quality.

We narrowed down on the following three pain points to inform our tool's design:

1. There is no existing standard review process or location, making review and collaboration especially difficult.
2. Users are oftentimes unsure about their decisions, leaving some questions partially or completely unreviewed.
3. Despite the process being inherently collaborative and involving multiple reviewers, users are left to their own to assign or review questions.

Prioritizing Collaboration, Efficiency, and Brevity

As mentioned earlier, a lot of the work reviewers currently have to do is tedious, error-prone, and inefficient. Therefore, in order to combat and mitigate some of these challenges, we decided to incorporate certain features into our prototype that were appropriate for the problem space.

List and Modular Views

The list view is intended to offer a faster and more streamlined approach to the review process only really emphasizing the basic functionalities such as accept or deny question and answer pairs. This configuration could allow reviewers to work their way through the questions faster, have a more holistic view of multiple questions at once, and would be more appropriate if the reviewers did not feel they needed to use more advanced features or functionalities.

The modular view offers more functionalities, and therefore more flexibility, for reviewing a question and answer pair. In the modular view, users can edit the questions to align them more to the intended result and enable collaboration which will be discussed in more detail below.

Approving, Denying, and Editing Functionalities

These were the basic functionalities that needed to be present for the synthetic data review process. A user, at the very least, needs to be able to approve a question if they believe it is accurate. Conversely, they should also be able to deny a question if they are dissatisfied with the synthetically generated data (SGD). However, if the user believes the question is not ready for either a denial or approval they should be able to provide their edits.

Collaborative Team Tooling

Collaborative team tooling essentially allows users to discuss the SGD with one another, allowing reviewers to see comments or feedback from other users. Additionally, we believed it would be beneficial to have the option to tag other users for any SGD in case a reviewer needs help or would like another perspective. We were also considering having

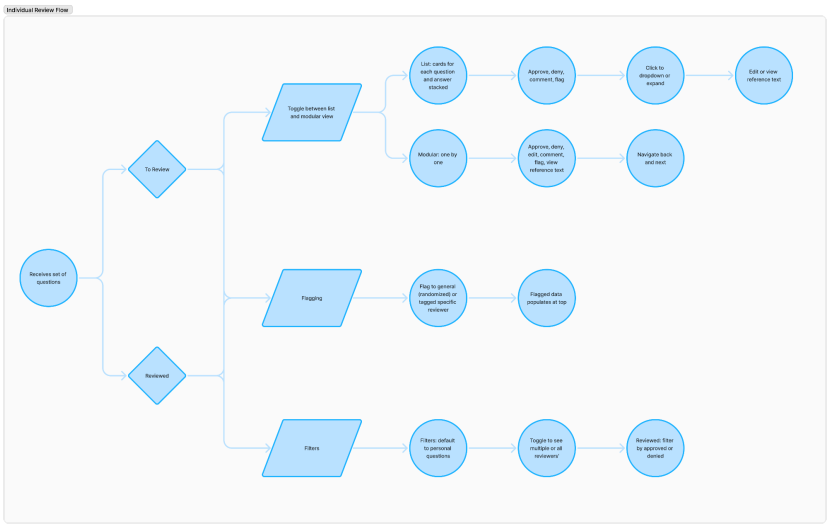
a hard assign option to hand over the question and answer pair to another review entirely if seen fit.

Reference Documents

While reviewing the questions, the reviewer should have access to the reference document the question is coming from, so they can best assess whether the question and answer pair makes sense.

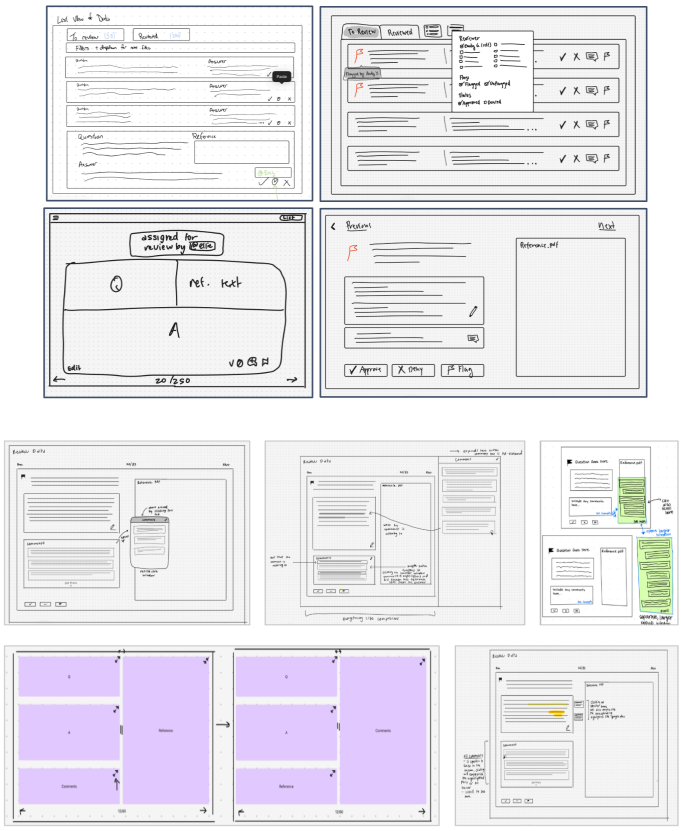
Designing Data Viewing and Interaction

In order to better understand how a set of data goes through the reviewal process, we mapped out the journey of how a user might navigate reviewing a set of Q&A's – from how they choose to view the data among larger sets to the final approval or denial stage.



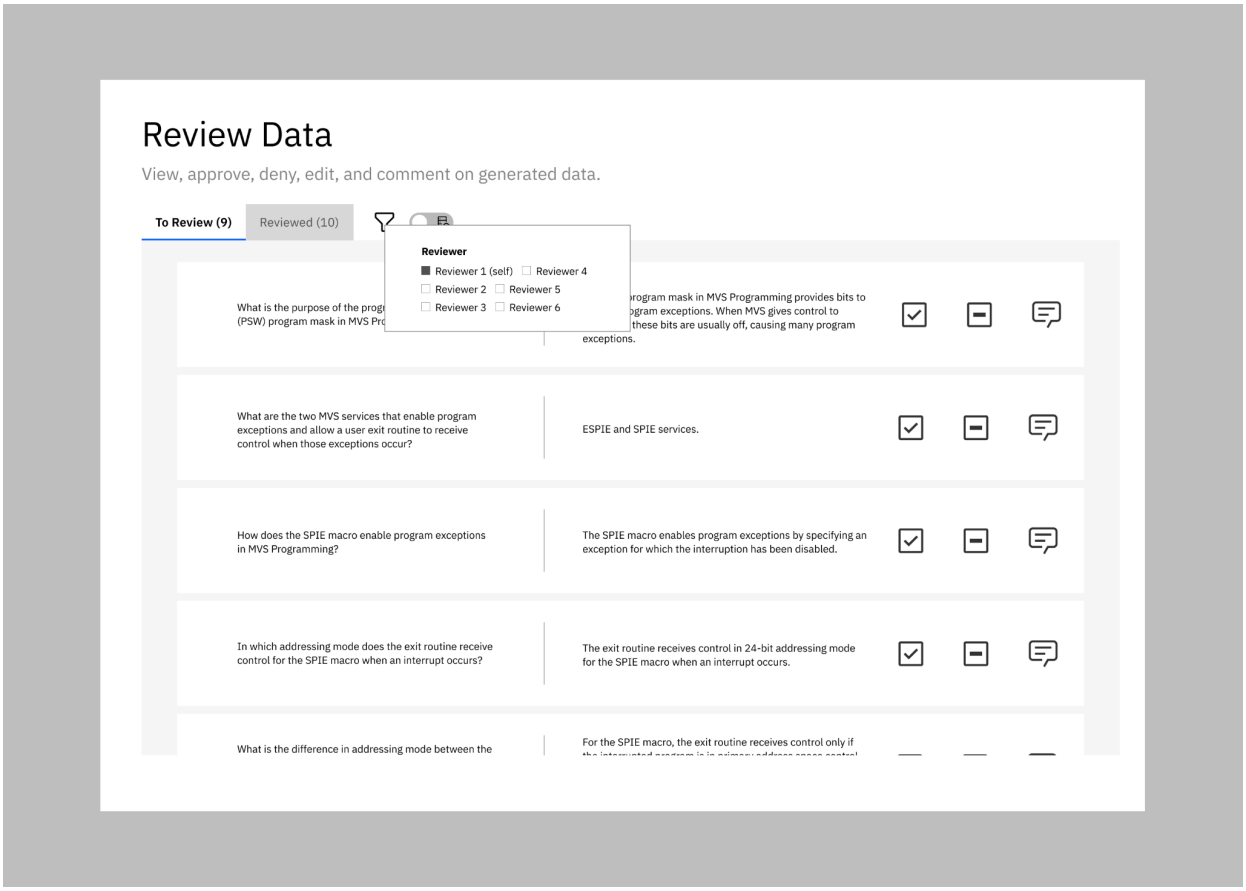
Collaboratively, we focused on generating multiple iterations of data viewing – how might a user quickly scroll through large sets of data with list view? How might a user focus on reference documents in a larger modular view?

Given our assumption of highly collaborative reviewing teams, we also spent time generating iterations of commenting, emphasizing the ability to reference the discussion during the review process and include tag other commenters in the process if a user was unsure about the content.

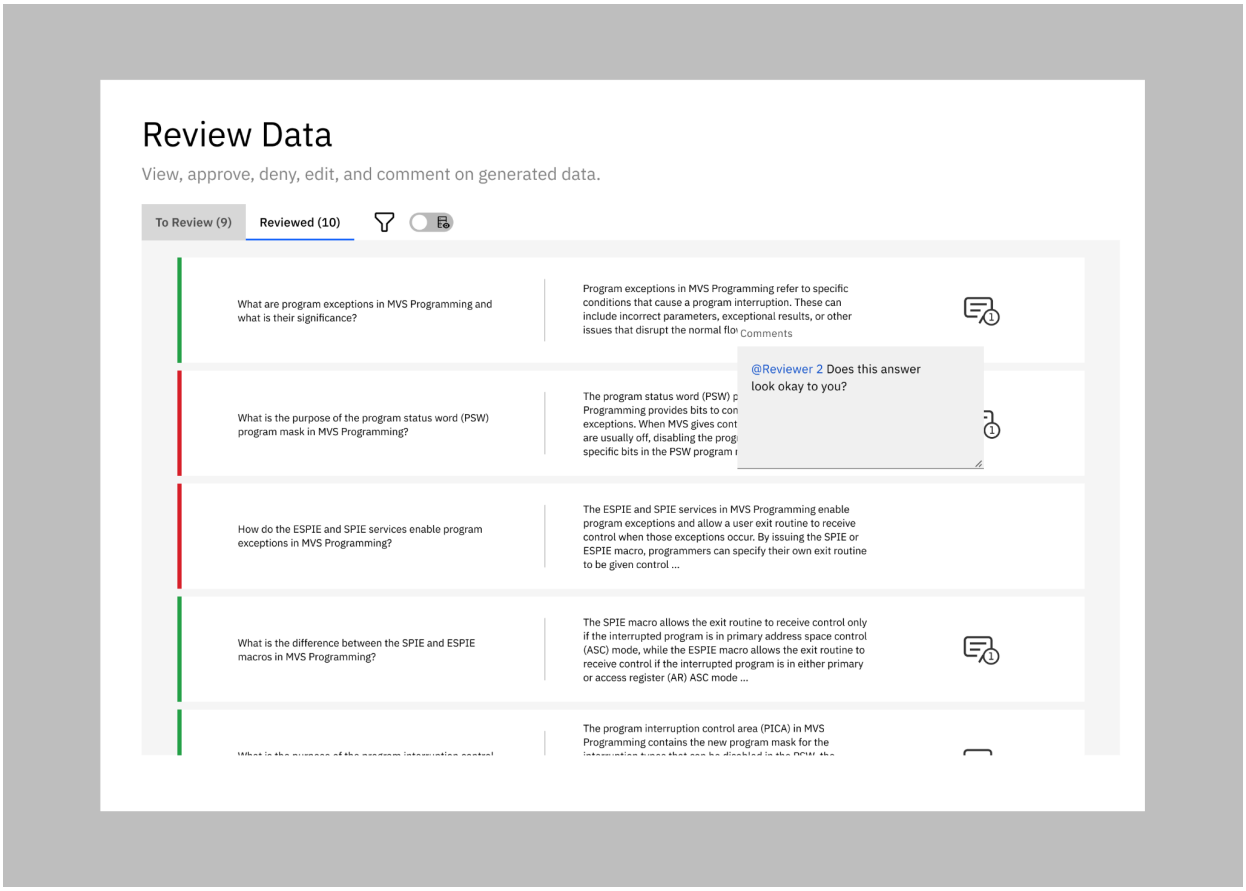


Using the current WatsonX interfaces as a reference and with continuous feedback from two WatsonX UX designers + one InstructLab developer, we designed screens encompassing the three prioritized features: Collaborative Team Tools (filtering, commenting), List and Modular Views, and Approving, Denying, Editing.

Users begin in list view, where they can easily see the content of the entire set of assigned Q&As. This view is mainly for users who spend a short amount of time on each of the questions, quickly approving or denying each set. This view is also for users who may want to check on the status of other reviewers, filtering through their team members' assigned Q&A.

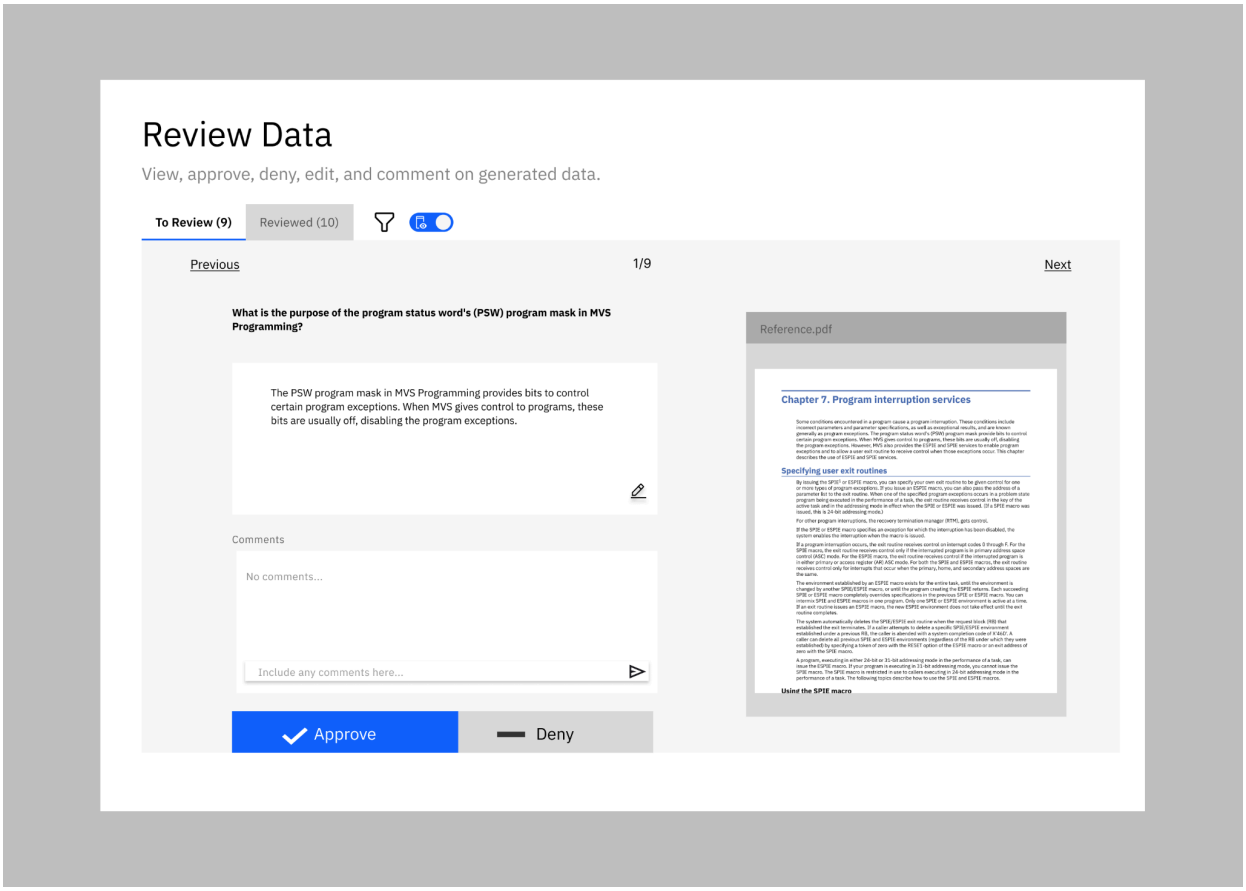


Users can also view their previously reviewed questions, easily able to see what's been approved or denied — and subsequently, the overall accuracy of the Q&As that have been generated by the model. However, the review process does not end upon the approval or denial of the question: reviewers can tag their team members in comments for a quick double-check, which will then appear in the tagged reviewer's "to review" feed.



Lastly, for users who are looking for a more comprehensive review and/or reference to the ground truth document, they can toggle “modular view”. Outside of the approve and denying functions, they’ll also see the option to edit the generated answer, providing a more accurate input for the model. They’ll also see a larger place to write comments if the question or answer warrants more in depth comments or discussion among team members.

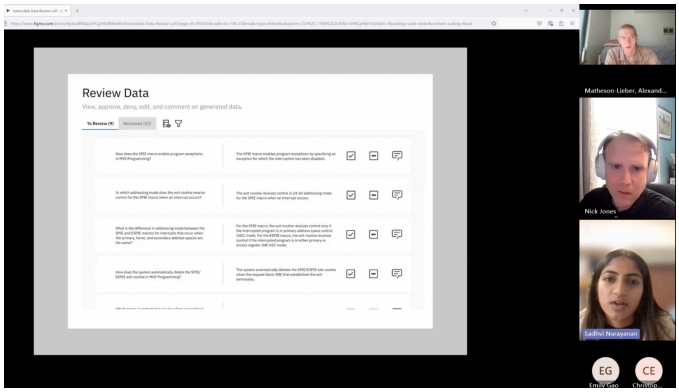
Users will be able to see the chunked reference document and scroll through it live to more accurately fact check the Q&A. From there, users can continue to previous and next q&as, and approve and deny as usual.

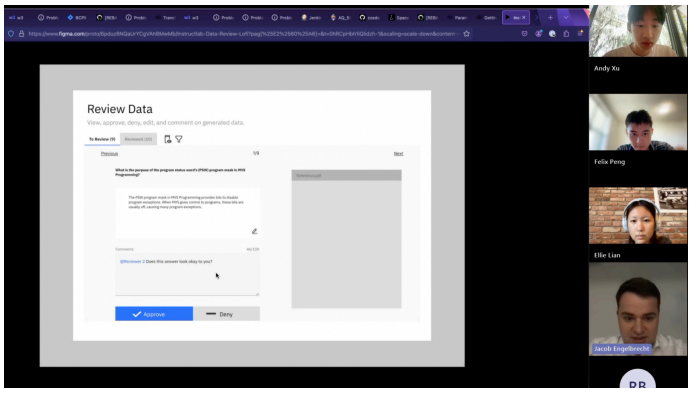


Testing with Developers [user research/testing]

With our mid-fidelity design completed, we turned to Software Developers at IBM who worked with the existing SDG process to better align our prototype with their current workflows. To do so, we conducted a user testing session with two of such developers.

Breaking into small groups, each with one developer and three designers, they walked through our prototype, sharing their thoughts and observations. Additionally, gathered further insights by asking questions on how they currently went about the review, how they used different features on our prototype, and what secondary features/edits they would like to see for a more intuitive platform.





Critical takeaways from this user testing session:

- Data review was a very individualized process
- Reviewers leaned heavily on the reference document
- Modular view was preferred to list view and more functional

Improving Navigation, Commenting, and Referencing

Completing user testing challenged our assumptions and gave us valuable insights into how to make our designs more intuitive and user friendly. After some reflection upon the feedback we had received, our team came up with a list of key pain points to prioritize solving and created a plan for our next design iteration. We decided to focus on making **navigation more intuitive, improving commenting, and increasing the accessibility of the reference document.**

Navigation changes

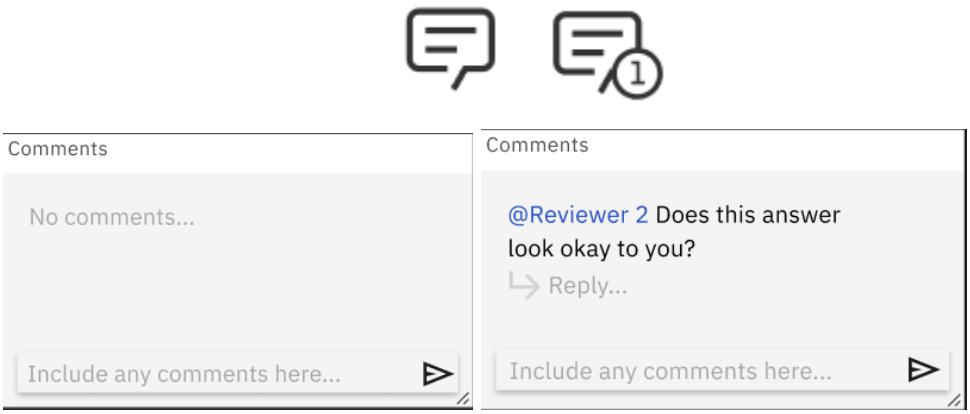
Within our testing, we noticed users had trouble toggling between the modular and list views. In order to address this issue, we redesigned the modular and list view toggle button to clearly show when a user was switching between sections and make navigating between sections more intuitive and clear.

Key changes:

- Switch icon is used to clearly indicate transition between sections
- List view icon to indicate current page
- Colorful animated transition to indicate the change to modular view

Commenting

From our testing, we found that the comment feature was not utilized as much as we had expected. Additionally, the developers that we worked with during testing made it clear that they prioritized a simple commenting experience and independent decision making, and that comment functions did not need to be overly complex (ex. Forums, etc.).

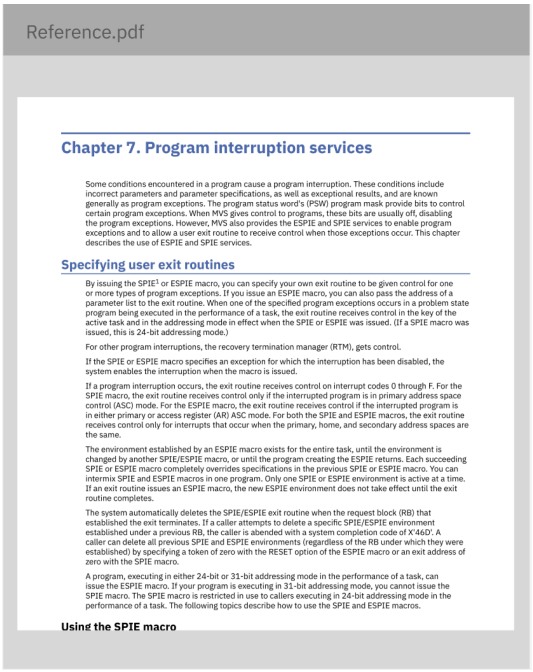


Key changes:

- Added minimally invasive comment display
- Redesigned comment modal to emphasize short interactions

Reference document

Within testing, we found that users referenced their own resource document for each question, and did not seem aware of the existing placeholder. Additionally users found it difficult to find the information they needed. We decided to add greater searchability and access to the reference PDF.

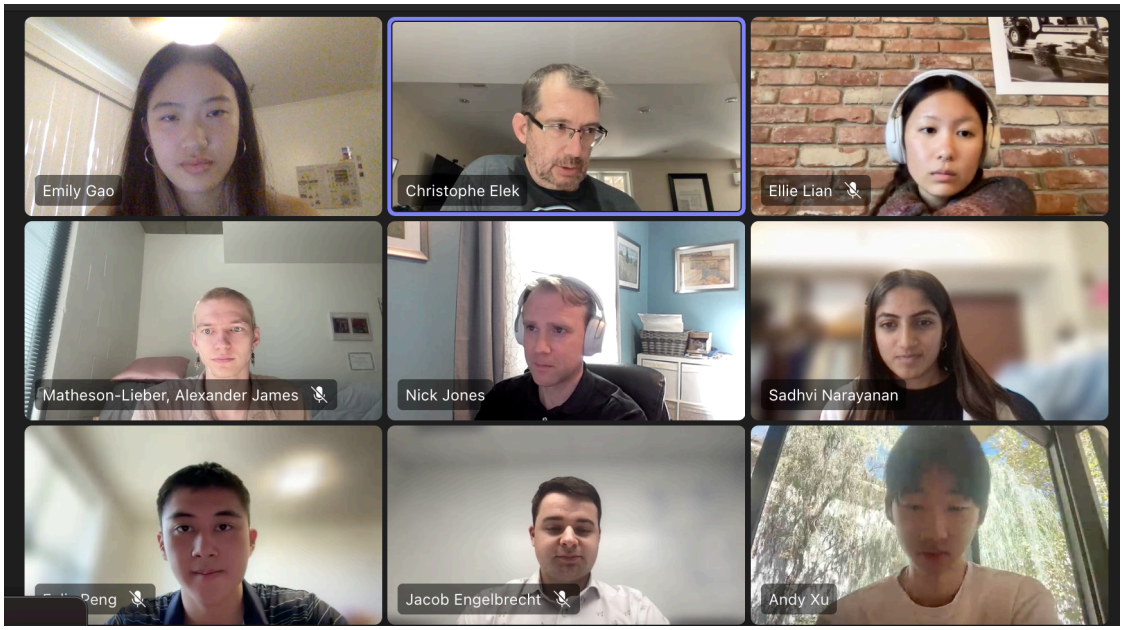


Key changes:

- Included reference doc for each question in modular view
- Have a PDF search tool
- Have reference document available in list view

IBM Stakeholder Presentation & Next Steps

The stakeholder presentation was a pivotal moment in aligning our design approach with IBM's strategic objectives for Watsonx AI. Key participants included senior representatives from UX, product management, and development teams. We showcased the design process as well as a walkthrough of the final prototype.



Stakeholders provided critical feedback, particularly on scaling the solution to meet real-world demands.

- Suhas (PM) emphasized the potential of extending the design beyond manual reviews to automated validation mechanisms.
- Nicholas (development) highlighted technical limitations like memory constraints and suggested considering more metrics or confidence scores alongside each dataset to assist with reviewing.

"This is a big step forward from what we've been doing in the past — a large Improvement!"

- Jacob Engelbrecht, Backend IBM SWE

Continued Design Exploration

There are a few secondary features that would push this design into a more advanced iteration:

- Dashboard for managers of reviewing teams to track reviewing statistics
- Metrics on the number of approved or denied datasets and confidence scores to evaluate overall synthetic data quality
- Role-based task assigning for teams to redirect questions to other members