Assisted Review Active Learning Guide

October 10, 2024 | Server 2022

For the most recent version of this document, visit our **documentation website**.

Table of Contents

1 Active Learning **5**

1.1 Installing the Active Learning application 5

1.2 Active Learning agents and instance settings 6

1.3 Basic Active Learning workflow 6

2 Creating an Active Learning project **9**

2.1 Creating the saved search 9

2.1.1 Optimizing your Active Learning project 9

2.2 Creating the classification index 11

2.3 Creating the reviewer group 12

2.4 Creating the review field 13

2.5 Creating the review layout 14

2.6 Creating a new Active Learning project 14

2.6.1 Fields 14

2.6.2 Active Learning objects 15

2.6.3 Viewing project settings and errors 16

2.7 Security permissions 17

2.7.1 Creating an Active Learning project 17

2.7.2 Viewing and editing an Active Learning project 17

2.7.3 Deleting an Active Learning project 18

2.7.4 Reviewer permissions 18

2.8 Reviewer access 19

2.8.1 Project admin considerations 19

2.8.2 Accessing the Active Learning review queue 19

2.8.3 Review workflow 21

3 Choosing an Active Learning review queue **22**

3.1 Prioritized Review 22

3.2 Coverage Review 22

3.3 Running Prioritized Review 23

3.3.1 Documents served in Prioritized Review 23

3.3.2 Family-based review considerations 23

3.3.3 Starting the Prioritized Review queue 25
3.3.4 Adding new documents 26
3.4 Running Coverage Review 27
3.4.1 Documents served in Coverage Review 27
3.4.2 Special considerations 27
3.4.3 Starting the Coverage Review queue 27
3.4.4 Adding new documents 28
4 Best practices for Active Learning review 30
4.1 General coding tips 30
4.2 Coding according to the "four corners" rule 30
4.2.1 Common scenarios which fail the "four corners" rule 30
4.3 Factors which affect Active Learning predictions 31
5 Monitoring an Active Learning project 32
5.1 Project Home dashboard 32
5.2 Update ranks 33
5.2.1 Running a search on a classification index 34
5.3 Document rank distribution 35
5.4 Prioritized review progress 37
5.5 Quality checks and checking for conflicts 37
5.5.1 Useful dashboards 37
5.5.2 Using these widgets 42
5.6 Determining when a project is done 42
6 Project Validation and Elusion Testing 44
6.1 Key definitions for Project Validation 44
6.2 Choosing a validation type 45
6.3 Starting Project Validation 45
6.4 Tracking sampled documents 46
6.5 Running Project Validation 47
6.5.1 Releasing unreviewed documents 47
6.6 Reviewing Project Validation results 48
6.7 Project Validation statistics 49
6.7.1 Defining elusion, richness, recall, and precision 50

6.7.2 How Project Validation metrics are calculated 51
6.7.3 How the Project Validation queue works 54
6.7.4 How the Project Validation sample is selected 54
6.7.5 How Project Validation handles skipped documents 55
7 Review statistics 56
7.1 Review Summary 56
7.1.1 Prioritized Review 56
7.1.2 Coverage Review 57
7.1.3 Project Validation 58
7.1.4 Manually-Selected 60
7.2 Model Updates 61
8 Active Learning performance baselines 62
8.1 Active Learning project size recommendations 62
8.2 Data set details 62
8.3 Classification index population + build results 62
8.4 Active Learning index build results 62
8.5 Update ranks results 63

1 Active Learning

Active Learning is a technology assisted review tool that helps you quickly organize your data and predict which documents are most likely to be relevant to reviewers. With very little training needed to see documents of interest, Active Learning can be used for cases of all sizes, even those as small as 1,000 documents. By using Active Learning, you can reduce the total time to review.

Active Learning works by using a technology called Support Vector Machine learning to continuously learn from your reviewers' coding decisions. Reviewers code documents using a binary classification system (for example, *Relevant* and *Not Relevant*). These coding decisions are ingested by the Active Learning model where machine learning takes place. As reviewers code, the model gets better at discerning what makes a document *Relevant* or *Not Relevant* and serves the best documents to reviewers. Active Learning provides two methods of review, making it flexible to your case needs:

- **Prioritized Review**—finds the documents most likely to be relevant to reviewers.
- **Coverage Review**—quickly separates your document into your two categories.

There are a number of tools available to help you monitor the progress of your review and eventually validate the success of Active Learning.

Note: For details on estimating recall in an Active Learning project, see Project Validation and Elusion Testing on page 44.

See these related pages:

- Creating an Active Learning project on page 9
- Choosing an Active Learning review queue on page 22
- Reviewer access on page 19
- Monitoring an Active Learning project on page 32
- Project Validation and Elusion Testing on page 44
- Security permissions on page 17
- Active Learning performance baselines on page 62

1.1 Installing the Active Learning application

To use Active Learning, you must first install the Active Learning application from the Application Library to your workspace.

Note: Once you install the Active Learning application, you can't uninstall it.

1. Navigate to a workspace where you want to install Active Learning.
2. Click the Relativity Applications tab.
3. Click New Relativity Application.
4. Next to **Application Type**, select **Select from Application Library**.
5. Click the ellipsis button in the **Choose from Application Library** field.

6. Select the Active Learning application, and then click **Ok**.
7. Click **Import**.
8. Review the import status of the application. Verify that the install was successful or resolve errors. For more information, see Installing applications in the Admin Guide.

1.2 Active Learning agents and instance settings

Ensure the following agents are installed and configured:

- Relativity Analytics Index Manager
- Analytics Index Progress Manager
- Analytics Categorization Manager
- Active Learning Manager (one per resource pool)
- Active Learning Worker

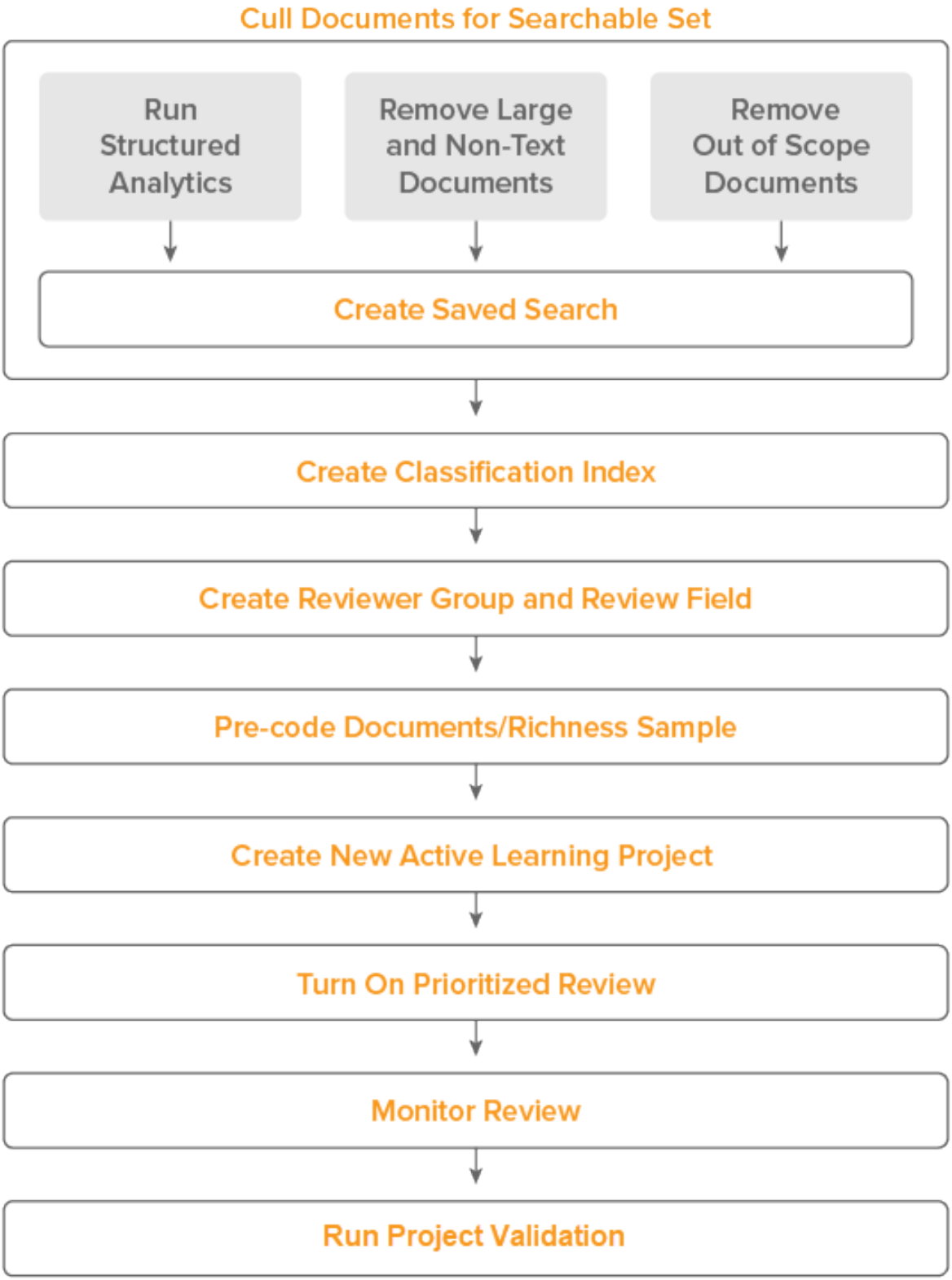
For more information, see the Agents Guide.

Active Learning uses the following instance settings:

- QueueBatchSize
- ReviewQueueRefreshThreshold
- ClassificationCategorizationDelay
- ClassificationCategorizationMaxDelay

1.3 Basic Active Learning workflow

The following graphic and corresponding steps depict a typical Active Learning workflow that integrates with other Analytics features. Note that each user's workflow may vary. You may not be required to follow all of these steps for every Active Learning project you run.


Cull Documents for Searchable Set
Run Structured Analytics
Remove Large and Non-Text Documents
Remove Out of Scope Documents
Create Saved Search
Create Classification Index
Create Reviewer Group and Review Field
Pre-code Documents/Richness Sample
Create New Active Learning Project
Turn On Prioritized Review
Monitor Review
Run Project Validation

1. Cull the documents you plan on using in the data source for your Active Learning project. This may include:
   - Running the following structured analytics operations:
     - Email Threading
     - Textual Near Duplicate Identification
     - (Optional) Language Identification
   - Removing large and non-text documents
   - Removing documents that are out-of-scope for your case
2. Create a saved search using the documents you identified.

   Note: Return only extracted text in the search.

3. Create an Analytics classification index using this saved search as the data source.
4. Create a review field with two choices available on the coding layout to be used by the reviewer group accessing the Active Learning project.
5. Pre-code documents in the data source on the review field by taking a richness sample. Aim for a balance of documents coded on the positive designation and the negative designation. Pre-coding documents by estimating richness helps speed up the Active Learning process and gives you a starting metric to gauge the progress of your Active Learning project later.
6. Create a new Active Learning project.
7. Turn on the Prioritized Review queue or Coverage Review queue, depending on your workflow. Have reviewers begin coding documents in the review queue.
8. Monitor the review queue.
9. Run Project Validation to validate the results of your project.

Note: For information on creating dashboards to help monitor your queue, see Quality checks and checking for conflicts on page 37.

2 Creating an Active Learning project

An Active Learning project consists of a few core components:

- **Saved search** - the set of documents to be used in the Active Learning project. This search is used to build the classification index.
- **Classification index** - the Analytics index which is the basis of the machine learning that takes place in Active Learning.
- **Reviewer group** - the group of reviewers coding documents in Active Learning.
- **Review field** - the field for reviewers to code documents in Active Learning. This field must be a single choice field with two choice

You must create each of these components before you can create an Active Learning project. Even if these items already exist in the workspace, you may want to create a new instance of each specifically for your new project.

2.1 Creating the saved search

The first component you must create for your Active Learning project is a saved search of all documents to be used in the Active Learning project. This set of documents is used as the data source when you create an Analytics classification index.

A few considerations for the search:

- The saved search must be public.
- The documents in the search must contain extracted text.
- For best results, we recommend no more than 9 million documents in the search.
- If you want to use family-based review (Prioritized Review only), entire families must be added to the saved search. This includes documents that you might otherwise exclude from an Active Learning project.

Note: If relevant documents are very rare, then targeted searches for relevant documents outside of the queue can help make Active Learning more effective.

2.1.1 Optimizing your Active Learning project

There are a few steps you can take when creating your saved search to optimize your Active Learning project.

2.1.1.1 Reduce document clutter

To streamline the start of your project, we recommend taking steps to reduce document clutter in your data source. This can include any combination of the following:

- Removing documents with poor or no text
- Removing out-of-scope documents

- Focusing on inclusive emails
- Running deduplication

2.1.1.2 Pre-code documents in the saved search

The Active Learning model only builds once you have at least five documents coded with the positive choice and five coded with the negative choice. We recommend having reviewers code a sample of documents prior to starting your Active Learning project to help start the model's learning and begin the project with a model build and a ranking of all documents. This can take place before or after you create the project.

Note: In Prioritized Review, if there are fewer than five coded documents for either choice, the system serves random documents until the threshold number of documents are coded. These documents appear under the Index Health column in Review Statistics.

There are a few ways to identify documents to pre-code, and you will usually combine more than one strategy:

- Draw a sample of documents with specific keywords, from key custodians, within a certain date range, etc. to help focus the sample on documents more likely to be important in the case. Repeat for different aspects of responsiveness to best train the model.
- Cluster your documents, then use the choice field stratified sampling script to draw an equal number of documents from each cluster. This ensures your sample will include a broad variety of documents. For more information, see Choice field stratified sampling in the Admin guide.
- Use pre-existing documents determined to be relevant/not relevant. You must ensure the coding decision for these documents is set on the review field used to create your Active Learning project.

Note: Aim for an even distribution of documents pre-coded on the positive/responsive designation as on the negative/not responsive designation when possible. For example, If your richness is low and you expect to find not responsive documents easily, then focus your effort on finding more responsive documents to pre-code.

2.1.1.3 Estimating richness

We recommend taking a richness sample to pre-code documents. Richness refers to the percentage of documents in a project that are responsive. We recommend calculating a richness estimate at the start of the project to give you a metric to gauge the progress of your project against.

To calculate richness, you can run an early Project Validation round which has Elusion with Recall selected. See Project Validation and Elusion Testing on page 44 for more information.

Alternately, you can use the following method:

First, use the Sampling tool to take a sample of documents from your data source and have reviewers code all of those documents on the project review field:

1. Navigate to the saved search used to create the Active Learning project.
2. Click the **Sampling** icon to display the sampling drop-down menu.
3. Select a **Sampling Type**, then determine a size or percentage appropriate for the size of your project. This does not need to be a large sample. Many people choose a Statistical sample using 95% Confidence Level and 5% Margin of Error, which is 385 documents or fewer.

4. Click **Sample**.
5. Have reviewers code the documents in the sample on the project review field. Do not skip documents. These pre-coded documents will help jump start your Active Learning project later.

To calculate the richness estimate, divide the number of documents the reviewers found to be relevant by the total number of documents in the sample.

- Sample Richness = # of documents coded on the positive choice / Total # of documents in sample

For example, imagine you ran a sample of 500 documents and your reviewers found 80 to be relevant. The richness for this sample would be 80 / 500 = .16, or 16%.

To estimate the number of responsive documents in the entire project, multiply the richness from the sample by the total number of documents in the saved search.

- Estimated # of responsive documents = Sample richness x Total # of documents in saved search

For example, if the total number of documents in the saved search used to create your Active Learning is 100,000, the estimate of the number of relevant documents for the entire project would be .16 x 100,000 = 16,000. This mean there will likely be around 16,000 relevant documents in your project. This prediction could be a little high or a little low, depending on the margin of error.

You can use this estimate to monitor the progress of your Active Learning project. For example, after you start your Active Learning project you see that your reviewers have coded 8,000 documents as relevant. Using the estimate, you can assume that you have found about half of all relevant documents in your case.

2.2 Creating the classification index

Active Learning uses an Analytics classification index as the basis of its machine learning. The classification index uses Support Vector Machine learning (SVM) to predict the relevance of uncoded documents in Active Learning based on their closeness to coded examples. To learn more about Support Vector Machine learning (SVM), see the Analytics Guide.

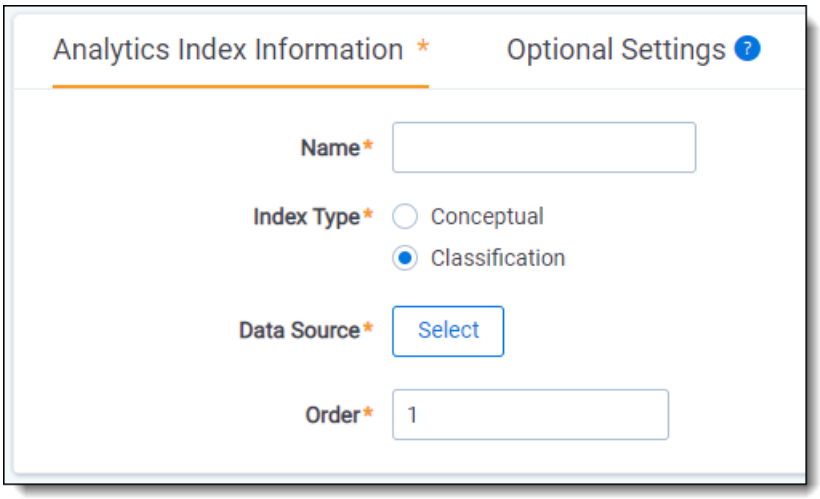
Once you have the set of documents you want to use for Active Learning in a saved search, you can create the classification index. To create the classification index:

Notes:

- Classification indexes are used only by Active Learning projects.
- You can add and remove documents from your Active Learning project (via the index) at any time.

1. Navigate to **Analytic Indexes**.
2. Click **New Analytics Index**.

3. Complete the following fields on the Analytics Index Information form:



   - **Name** - enter a name for the index. This value appears in the search drop-down menu after you activate the index.
   - **Index type** - select **Classification**.
   - **Data source** - select the saved search created above.
   - **Email notification recipients** (under Optional Settings) - send email notifications when your index successfully completes, fails, or when the index is disabled because it is not being used. Enter the email address(es) of the recipient(s). Use a semicolon to separate multiple email addresses. A message is sent for both automated and manual index building.
   - **(Advanced Settings fields) Relativity Analytics server** - select the Analytics server that should be associated with the Analytics index. This list includes only Analytics servers added to the resource pool for the workspace.
4. Click **Save**. Once you click Save, the Analytics Index console becomes available.
5. From the Analytics Index console, click **Run**. If this is your first time running the index, it will automatically run a full population. Otherwise, choose **Run**, then **Full**.

You can continue creating the other required components for Active Learning while you wait for the index to activate. However, the index must be active in order to create the Active Learning project.

2.3 Creating the reviewer group

When you create an Active Learning project, you designate a reviewer group which is the set of users you want to code documents in the Active Learning project. You can create a new group specifically for your Active Learning project, or you can use an existing group. Ensure the group is added to the workspace where you're running Active Learning.

Notes:

- The users in this group are not automatically added to the Active Learning project. You must grant each individual access to the Review queue.
- You can add or remove users from the group at any time.

The reviewer group accessing the Active Learning project must have the following permissions:

Object Security	Tab Visibility	Other Settings
▪ **Active Learning Reviewer** - View ▪ **Document** - View, Edit	▪ Documents	▪ **Browsers** - None OR Folders and/or Field Tree and/or Clusters

- If Browser permissions are set only to Advanced & Saved Searches, reviewers can't access the view for the review queues.
- The review group must have access to all documents in the Active Learning project. If documents are not accessible (for example, documents stored in a secured folder) they are not served to reviewers in the queue.

2.4 Creating the review field

Active Learning requires a single choice field with two choices for reviewers to code on. One choice must represent the positive/responsive designation, and the other the negative/not responsive designation. Reviewers may code documents on other fields that you include in your review layout; however, only this review field is used by the Active Learning model.

1. In your workspace, navigate to the **Fields** tab.
2. Click **New Field**.
3. Complete the fields in the **Field Information** section as follows:
   - **Name** - enter a name for your review field.
   - **Object Type** - select **Document**.
   - **Field Type** - select **Single Choice**.
4. Under **List and Dashboard Settings**, select **Enable Group By** and **Enable Pivot**.
5. Leave all other fields as default, or edit to your workspace specifications.

   Note: We recommend turning off family propagation with Active Learning.

6. Click **Save**.
7. In the **Choices** section, click **Add Choice**.
8. Enter a name for the positive/responsive designation, and then click Enter.
9. Click **Add Choice** again.
10. Enter a name for the negative/not responsive designation, and then click Enter.

Note: You can edit these choices after you create the project.

2.5 Creating the review layout

Active Learning does not create a layout automatically; you must create or modify an existing layout for reviewers to make coding decisions using the review field created above. This layout is not a prerequisite for creating your Active Learning project, but the layout should exist before the reviewer group begins coding in Active Learning. You can also include other fields on this layout. Active Learning only learns from the review field.

2.6 Creating a new Active Learning project

Once you've configured the required workspace components, you can create your Active Learning project

To create an Active Learning project, complete the following:

1. From your workspace, navigate to **Assisted Review** > **Active Learning Projects**.
2. Click **New Active Learning Project**.
3. Complete the fields on the layout. See Fields below.
4. Click **Save**.

You are then redirected to the project home dashboard.

Note: Analytics classification indexes are copied over when a workspace is used as a template, as are conceptual indexes. However, you can't copy an Active Learning project in a template.

2.6.1 Fields

The Active Learning Project layout contains the following fields:

- **Project Name** - enter a name for your Active Learning project.
- **Analytics Index (Classification)** - select the Analytics classification index. The index must have completed a build and cannot be associated with any other Active Learning project.
- **Review Field** - select the single choice field you created for review. This field must have exactly two choices. You can edit these choices after project creation.
- **Positive Choice** - select the choice that represents the positive/responsive designation.
- **Suppress Duplicate Documents** - selecting this option stops Active Learning from serving documents that are textually similar to other coded documents to be coded (this does not consider word order). This option reduces the total number of documents needed to be reviewed. Since these are not exact duplicates, you will most likely need to review suppressed documents after the project is complete. For more information, see Suppressed duplicate documents on the next page.

  Note: We recommend setting the Suppress Duplicate Documents setting to No for Prioritized Review and Yes for Coverage Review. Note that you cannot change this setting once you create your project.

- **Reviewer Group** - select the group you want to access the Active Learning queues.

Note: Once a project is created, you cannot edit the fields.

2.6.2 Active Learning objects

Upon saving a new project, Relativity creates a few different objects specific to Active Learning.

2.6.2.1 Project document list view

This view is automatically secured to the reviewer group. Initially the view will be empty, but as reviewers code documents in Active Learning queues, the view returns documents previously reviewed by the currently logged in reviewers. This is the document list view that's tied to the Active Learning project. It has the same name as the Active Learning project and is customizable by admins. This is the only place a reviewer can enter a project queue.

2.6.2.2 Project fields

Relativity also creates new fields that can be used for custom document list reporting. These fields are updated per document as reviewers code on the project review field.

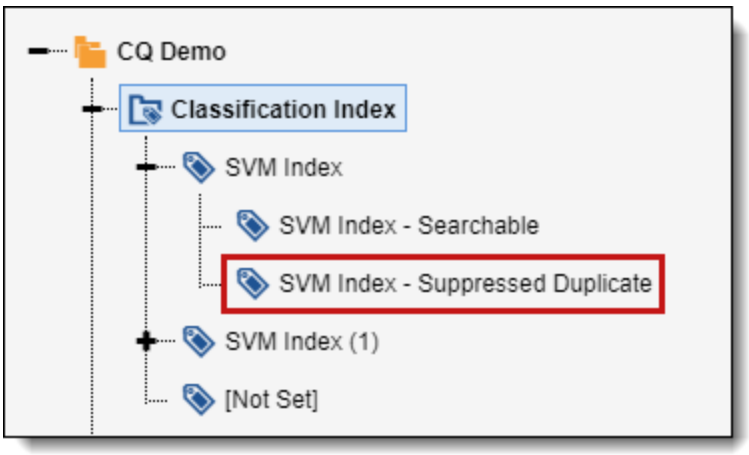
- **<Project Name> Reviewers::User** - the name of the reviewer who coded the document in the queue.
  - You can use this field to identify manually-selected documents. Search for documents where the review field **is set**, but this field **is not set**. Because manually selected documents are documents which are not coded via an Active Learning queue, the Reviewers field would not be set for those documents.
- **<Project Name> Reviewers** - the name of the project plus the name of the reviewer who coded the document in the queue.
- **<Project Name> Reviewed On** - a timestamp indicating when a document was reviewed.
- **<Project Name>:: Coverage Review** - the 200-document interval in which the document was reviewed in the Coverage Review.
- **<Project Name>:: Project Validation** - the Project Validation in which the document was reviewed.
- **<Project Name>:: Prioritized Review** - the 200-document interval in which the document was reviewed in the Prioritized Review.

For more information on using these fields in reports and dashboards, see Monitoring an Active Learning project on page 32.

2.6.2.3 Suppressed duplicate documents

If you chose to suppress duplicate documents, you can view those documents from the Field Tree. You can find the tag for suppressed duplicate documents under the Classification index associated with the Active Learning project. Documents are tagged with the **<Index Name> - Suppressed Duplicate** tag as they're identified during review.

There is currently no way to add these documents back into the review queue or to identify which coded document caused a document to be suppressed. However, you can code these documents manually and they are indicated as manually-selected documents in the project.



2.6.3 Viewing project settings and errors

Once you've created an Active Learning project, you're redirected to the Project Home tab. In the top-right corner of the Project Home tab are three icons.

- **Update Ranks** - use to manually update the document ranks and ensure the rank categorization field is up to date. For more information, see Update ranks on page 33.
- **Notifications** - informs you of any notifications you may have in the project such as project errors . A red circle appears in the bottom corner of the Notifications icon if a notification exists. Follow the instructions in the error message to resolve the project error. Once you've resolved the error, the error message automatically disappears.
- **Project Settings** - gives an overview of the settings that were either selected during the initial project setup, or that can be modified mid-project. It includes the following:
  - Project Name
  - Analytics Index
  - Review Field
  - Positive Choice
  - Suppress Exact Duplicates
  - <Positive Choice> Cutoff

    Note: Click the icon to edit the <Positive Choice> Cutoff value. This rank divides positive and negative predictions, such as "relevant" and "not relevant." When you update the cutoff value, the value is updated in all three places where it's used in the application: Project Validation, Update Ranks, and Project Settings.
  - Reviewer Permission Group

2.7 Security permissions

This page contains information on the security permissions required for creating and interacting with an Active Learning project.

2.7.1 Creating an Active Learning project

To create an Active Learning project, you need the following permissions:

Object Security	Tab Visibility
▪ **Active Learning Project** - View, Edit, Add ▪ **Active Learning Review** - View, Edit, Add, ▪ **Active Learning Reviewer** - View, Edit, Add ▪ **Analytics Categorization Result** - View, Add ▪ **Analytics Categorization Set** - View, Edit, Add ▪ **Analytics Categorization Set Build History** - View ▪ **Analytics Index** - View, Edit ▪ **Field** - View, Edit, Add ▪ **Object Rule** - View, Add ▪ **View** - View, Add, Edit Security ▪ **Workspace** - Edit Security	▪ Active Learning Projects

2.7.2 Viewing and editing an Active Learning project

To view and edit an Active Learning project, you need the following permissions:

Note: It's not possible to have "view-only" Active Learning permissions. If you have permission to view but not edit an Active Learning project, you aren't able to view the project.

Object Security	Tab Visibility
▪ **Active Learning Project** - View, Edit ▪ **Active Learning Review** - View, Edit, Add ▪ **Active Learning Reviewer** - View, Edit ▪ **Analytics Categorization Set** - View, Edit ▪ **Analytics Categorization Set Build History** - View ▪ **Analytics Index** - View	▪ Active Learning Projects

2.7.3 Deleting an Active Learning project

In order to delete an Active Learning project, you need either workspace admin or system admin permissions. This will allow you to delete the view created by the project.

2.7.4 Reviewer permissions

The reviewer group accessing the Active Learning project must have the following permissions:

Object Security	Tab Visibility	Other Settings
▪ **Active Learning Reviewer** - View ▪ **Document** - View, Edit	▪ Documents	▪ **Browsers** - None OR Folders and/or Field Tree and/or Clusters

- If Browser permissions are set only to Advanced & Saved Searches, reviewers can't access the view for the review queues.
- The review group must have access to all documents in the Active Learning project. If documents are not accessible (for example, documents stored in a secured folder) they are not served to reviewers in the queue.

Note: The users in this group are not automatically added to the Active Learning project. You must grant each individual access to the Review queue.

2.7.4.1 Resolving reviewer permissions warnings

If the reviewer group accessing an Active Learning project does not have the required permissions to review the documents in that project, a warning will appear in the notifications area in the top-right corner of the Active Learning Project Home tab. To view the warning, click the notification icon.

The permissions warnings can be resolved as follows:

Permissions Warning Message	Solution
User Group "Group Name" does not have Reviewer Object **view** permission.	Grant the user group **View** permissions on the Active Learning Reviewer Object. See Setting workspace permissions on the Relativity Documentation site.
User Group "Group Name" does not have Document Object **edit** permission.	Grant the user group **Edit** permissions on the Document Object. See Setting workspace permissions on the Relativity Documentation site.
User Group "Group Name" may not have Document item level security or other security permissions.	Grant the user group **Edit** permissions to any documents within the Active Learning project which are individually secured or held in private folders. To adjust access to individual documents or folders, see Relativity object security on the Relativity Documentation site.

2.8 Reviewer access

After a project admin creates and saves a new Active Learning project, Relativity creates a new document list view that's tied to the Active Learning project. This view is automatically secured to the reviewers added to the project. As an Active Learning reviewer, this view is the only place you can enter the review queue.

Before assigning a document to a reviewer, Relativity ensures the following conditions are met:

- The current reviewer has permission to the document.
- The document has not been coded already.
- The document is not assigned to any other active reviewer.

Documents can only be assigned to one reviewer except in cases where the reviewer leaves the viewer without making a coding decision or is automatically logged out. When this happens, the queue can reassign the document to another reviewer. Once a document has been coded by a reviewer, the **Project Reviewers::User field** will show that reviewer's name.

Reviewers code documents on the review field specified for the project and then click **Save and Next** to get a new document. They can also code other fields not associated with the project.

2.8.1 Project admin considerations

Note: We recommend waiting until non-business hours to update security permissions for document folders containing a large number of documents. Doing the update during business hours could lead to your workspace not loading and a delay in your document review project.

- If you don't want reviewers to skip documents, make the project review field a required field.
- We recommend turning off family propagation with Active Learning.
- Make sure all reviewers assigned to the project are enabled users in the Relativity environment. For more information, see Creating and editing a user on the RelativityOne documentation site.
- You can grant permissions to specific users or groups for certain icons in the Viewer. Any permissions changes will affect how those groups see the Viewer for both Active Learning and non-Active Learning projects.
- Reviewers can change the coding decision on documents they previously reviewed. These documents aren't considered manually-selected documents. The next model build will include the most recent coding update.

2.8.2 Accessing the Active Learning review queue

To enter the Active Learning review queue, navigate to the Documents tab and ensure you are on the view named after the Active Learning project. This document view returns a list of all documents previously reviewed by you in the queue. When you first access the view, no documents appear because none have been coded. As you code documents from the Active Learning queue, documents appear in the list.

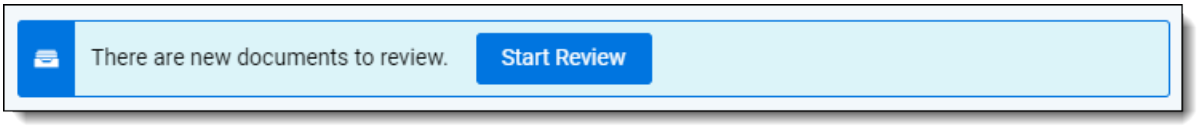
The project document view enables or disables reviewer access based on whether you are added to the queue and whether the queue is active.

If you are added to the queue and the queue is enabled, a blue banner appears in the project document view with a **Start Review** button. Clicking **Start Review** begins the review process.

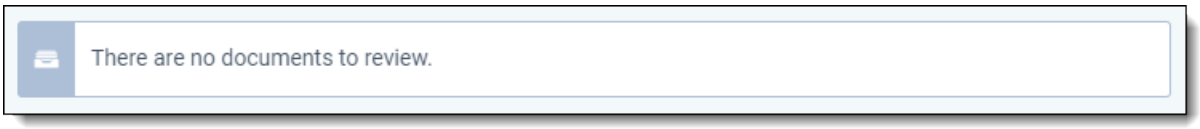
If the Start Review button is not appearing as expected, have the project admin try these troubleshooting steps:

1. Ensure reviewers are added to the group associated with the Active Learning project.
2. Ensure the reviewers have the proper security permissions. For more information, see Reviewer permissions on page 18.
3. Ensure the saved search associated with the Active Learning project contains documents with extracted text.
4. Ensure the classification index associated with the Active Learning project is active.
5. On the Project Home tab, ensure the **Start Review** button is selected.
6. When previewing permissions, ensure that you are previewing user permissions, not group permissions. The Start Review button does not appear in permissions group previews.
7. Ensure the following agents are installed and running:
   - Active Learning Manager
   - Active Learning Worker
   - Analytics Categorization Manager
8. In SQL, ensure there aren't orphaned jobs in the **EDDS.ActiveLearningProjects table** that don't correspond with an existing project. In other words, you look for projects that were deleted on the front end but still exist in the EDDS.ActiveLearningProjects table.
9. Check the **EDDS.AnalyticsCategorizationSetQueue** table and look for any 'ERROR' jobs.
10. Ensure rows exist (records) in the **EDDSDBO.DRQ_<long guid>** table in the workspace database. If there are no records here that means documents are not being properly submitted to the Active Learning Review Queue.
11. Ensure the language is English on the agent server where the Active Learning agents reside.

Note: If the queue is disabled or you're removed from the queue while reviewing documents, a warning message will appear in the viewer.



If you aren't added to the queue or if the queue is paused, you can't access the Active Learning queue from the project document view. You can still access the view itself but only for reviewing previously coded documents. You may still code documents outside the queue if you have permissions to do so.



2.8.3 Review workflow

Click **Start Review** to access the Active Learning queue viewer and start reviewing documents. The Viewer looks the same no matter which type of queue is active, contains the same functionality available for a standard queue, and relies on the same permissions.

Code documents on the review field specified for the project and then click **Save and Next** to get a new document. You can also code other fields not associated with the project as determined by your project admin.

Reviewers must code documents based on the so-called "four corners" rule. This means that a document should be judged relevant or not relevant based solely on the extracted text of that document only. Documents that are relevant based on family members should not be coded relevant on the Active Learning review field. Although individual anomalies will not hurt the algorithm, too many in aggregate could cause the model to perform worse.

2.8.3.1 Skipping documents

You can skip documents without making a coding decision unless the review field is a required field. You can also click to skip the current document and open a new document. In the upper right of the Viewer, you can see a running count of all documents coded in the project, including skipped documents. Skipped documents are also reflected in the Project Home dashboard.

Note: Reviewers can change the coding decision on documents they previously reviewed. These documents aren't considered manually-selected documents. The next model build will include the most recent coding update.

3 Choosing an Active Learning review queue

Once you've created an Active Learning project, you're redirected to the Project Home tab. From here, you can view a dashboard with information about your project and select an Active Learning review queue to start your review. For more information about the Project Home dashboard, see Monitoring an Active Learning project on page 32.

There are two types of review queues in Active Learning: Prioritized Review and Coverage Review. An Active Learning workflow starts in either the Prioritized Review or Coverage Review queue, moving to the Project Validation queue when the project is nearing completion. See below for a description of the review queues.

3.1 Prioritized Review

The Prioritized Review queue serves the highest ranking documents, in other words, documents that are most likely to be coded on the positive choice (ex. Responsive). The goal of Prioritized Review is to quickly identify and review the most relevant documents in your set. If you are unsure which queue to select, we recommend using the Prioritized Review queue.

When coding documents in the Prioritized Review queue, a small set of documents are included for index health. The documents included for index health are system-selected to help the model develop a better range of training.

Use the Prioritized Review queue if:

- You need to quickly locate and review the most relevant (responsive) documents in your set.
  - For example, a standard request for production with a need to do redaction and/or privilege review.
- You want to review relevant documents and their family together.
- You have a document set with lower richness. For information on estimating richness, see Creating an Active Learning project.
- You are unsure which queue to select.

For more information, see Running Prioritized Review on the next page.

3.2 Coverage Review

Coverage Review serves up documents that are optimal for training the model. The goal of Coverage Review is to quickly separate your documents into the positive choice and negative choice categories. Unlike Prioritized Review, which serves up the highest ranking documents, Coverage Review is intended for quick production use cases where you want to train the model as fast as possible.

The documents that are served during Coverage Review can be either relevant or non-relevant and are chosen by the engine solely to train the model. Coverage Review serves up the documents the model is most unsure about. Coverage Review continues serving up documents until there are no longer any documents to review. However, you should stop reviewing documents much earlier, once the model has stabilized.

Use the Coverage Review queue if:

- You need to quickly classify your documents into relevant/not relevant sets.
  - For example, to respond to HSR second request or if you have an incoming production.
- You have a large case and don't necessarily need to review and code every relevant document. For example:
  - A government production request with a tight timeline.
  - An internal investigation where you are mining for information for your own benefit.

Note: Coverage Review works best with document sets with greater richness (>2%).

For more information, see Running Coverage Review on page 27.

3.3 Running Prioritized Review

The Prioritized Review queue serves documents that are most likely to be coded on the positive choice (ex. Relevant) with a small set of documents included for index health. The documents included for index health are system-selected to help the model develop a better range of training. Prioritized Review has the option of turning on family review, which lets you leverage the efficiency of Active Learning while letting you review and code all documents in a family together.

Note: Family-based review is optimized for email and attachments. Textual near duplicates and threads aren't supported.

3.3.1 Documents served in Prioritized Review

The Prioritized Review queue serves up a mixture of documents: 10% of documents are randomly selected; 20% of documents are chosen for scores "in the middle" of the review (in the 40 to 60 range) for index health; and the final 70% are the highest-ranking uncoded documents remaining in the project. This means for every 200 documents, around 140 (70%) are chosen for being highly ranked. If you wanted to calculate the relevance rate manually, you would take **Number of Highest Ranked Coded [Positive Choice] / Number of Highest Ranked**. You can view these values in the Prioritized Review queue review summary on the Review Statistics tab. For more information, see Review Statistics.

Note: If you start the Prioritized Review queue before a model build occurs, the queue initially serves up random documents because no scores are available for documents until a model build has completed.

3.3.2 Family-based review considerations

Prioritized Review gives you the option to review family documents (optimized for email and attachments) together. When you enable **Include Family**, Active Learning serves documents with their family documents to reviewers. All family documents included in the index are served to the reviewer, including documents that were previously coded.

There are several special considerations to keep in mind as this setting is not needed for every case. Review the following special considerations prior to starting your review.

3.3.2.1 Project setup considerations

In order for family documents to be served together in Prioritized Review, the parent documents and their family must all be added to the classification index. This includes documents without text.

Note: Textually empty documents will receive a rank at 50 (visible on the histogram), but they must be included in order to be served up to reviewers.

3.3.2.2 Prioritized Review considerations

- Family-based review is only available for new projects or Prioritized Review queues that have never been started.
- Once you click **Start Review**, you can't edit the **Include Family** settings. This means you can't turn family-based review on or off mid-project.
- If you have any families with more than 100 documents, we recommend removing those families from the classification index and reviewing them manually. Large families can slow down the review process.

Reviewer considerations

- Family review is optimized for email and attachments, and documents are sorted in the queue by control number. Using a relational field other than the family group relational field, such as the email thread group or textual near duplicate identification group fields can cause undesirable results.
- When family-based review is enabled and the document served to a reviewer is part of a family, the documents are presented to the reviewer in ascending control number order.
  - Upon requesting a document, the document with the lowest control number is served first, which is typically the parent document of the family
  - When a reviewer clicks Save and Next, the next document in the family is served, letting the reviewer review family documents in an intuitive order.
  - This means the first document you see may not necessarily be the document most likely to be relevant; you could see a family document first.
  - Note that index health and randomly selected documents are also served with their families when family-based review is enabled.
- Reviewers must code documents based on the so-called "four corners" rule. This means that a document should be judged relevant or not relevant based solely on the extracted text of that document only. Documents that are relevant based on family members should not be coded relevant on the Active Learning review field. Although individual anomalies will not hurt the algorithm, too many in aggregate could cause the model to perform worse.
- A document and its family are assigned to a single reviewer. If a reviewer is removed from the project after a set of family documents have been assigned and reviewed, you may need to manually find those unreviewed "abandoned" family documents.

  Use the following saved searches to find all unreviewed "abandoned" family documents:
  Search 1

  - Conditions: **<Project Name> Reviewed On** is set
  - Related items:
    - + Family

- Fields:
  - <Review field>
  - <Project Name>::Prioritized Review
  - <Project Name> Reviewed On
  - <Project Name> Reviewers::User
  - <Family Field>

Search 2

- Conditions: **(Saved Search)** document is in **<Saved Search 1>**
  AND
  <Project Name> Reviewed On is not set

  Note: To find unreviewed documents for a specific user, **<Project Name> Reviewers::User** equals **<specific user>** in place of **<Project Name> Reviewed On** is not set.

- Fields:
  - <Family Field>

3.3.3 Starting the Prioritized Review queue

To start the Prioritized Review queue:

1. Add reviewers to the queue.

   Note: We recommend no more than 150 concurrent reviewers per project. Concurrent reviewers are defined as reviewers making coding decisions in an Active Learning queue at the same time. There is no limit to how many reviewers you can add to a queue as long as the number of concurrent reviewers remains at 150 or fewer.

   a. Click **Add Reviewers** in the queue modal.
   b. Find and select an individual reviewer, or type to enter a reviewer name. You can also click **All Users** to select every user in the group.
   c. Click the green check mark to save your changes.
   d. Click the red X to cancel reviewer changes.

2. If you want to review family documents, toggle the **Include Family** field. For more information, see Family-based review considerations on page 23.

   Note: Once you click **Start Review**, you can't change your family review selection.

   a. Next to **Family Field**, select the relational field you want to review from the drop-down. This drop-down includes the friendly name of all relational fields in the workspace. Note that family review is optimized for email and attachments, which is typically called Group Identifier or Family Identifier.

3. Click **Start Prioritized Review**.

The next time an active reviewer accesses the view for the Active Learning project, they see a blue banner which indicates they have access to the queue. For more information, see Reviewer access on page 19.

You can see the following items in the Prioritized Review queue modal:

- **Coded** - the number of documents coded and skipped in the Prioritized Review queue.
- **Active Reviewers** - the number of reviewers added to the project.

Admins can pause the review at any point in the project by clicking the **Pause Prioritized Review** button on the queue modal. Once the review is paused, the access point on the document view is disabled. You can restart the review at any time. Documents coded on the review field while the review is paused are included the next time the model rebuilds.

3.3.4 Adding new documents

If you want to add or remove documents from your Active Learning project after review has begun, complete the following:

1. Click **Pause Review** to stop the review queue.
2. Add or remove documents to the saved search used as the data source for the classification index. Be sure to return only the extracted text in your saved search.
3. Navigate to the **Analytics indexes** tab, and then click the classification index used to create your Active Learning project.
4. On the index console, click **Run**, then **Incremental**.
5. Once the index finishes populating, return to your review queue and click **Start Review**.

Notes:

- All documents, including the newly added documents, are given a rank score after the incremental population.
- The following explains the way Active Learning handles documents removed from the project:
  - Documents in a queue that have been coded and then deleted from the workspace are marked as skipped. If the document was manually coded, then deleted, the manually coded statistics update to the correct values.
  - When the index is repopulated, deleted documents and documents removed from the search are removed from the model. After repopulation, the Project Size statistics and the count of coded documents are updated.
  - If Project Validation is in progress when an index is repopulated, the Project Validation is canceled.

If one of the new documents is part of a group of family documents that was already reviewed in Active Learning, the document can be reviewed by any reviewer. The other previously reviewed family documents won't be served again with this new document.

3.4 Running Coverage Review

Coverage Review serves up documents that are optimal for training the model. The goal of Coverage Review is to quickly separate your documents into the positive choice and negative choice categories. Unlike Prioritized Review, which serves up the highest ranking documents, Coverage Review is intended for quick production use cases where you want to train the model as fast as possible.

The documents that are served up during Coverage Review can be either relevant or non-relevant and are the most impactful to training the model. Coverage Review begins by serving the documents the model is most unsure about - these are documents with a rank near 50. Coverage Review continues serving up documents until there are no longer any documents to review. However, you should stop reviewing documents much earlier, once the model has stabilized.

3.4.1 Documents served in Coverage Review

The Coverage Review queue serves up a mixture of documents: 45% are documents with ranks immediately above 50, 45% are documents with ranks immediately below 50, and the remaining 10% are documents with a rank of exactly 50.

For the documents above and below 50, the queue serves up the ranks closest to 50 before serving up ranks that are further away. For example, all documents ranked at 49 are served up before documents ranked at 48, documents ranked at 51 are served up before documents ranked at 52, and so on. This applies human coding decisions to the most ambiguous documents, training the model and improving its predictions about the remainder.

Note: If you start the Coverage Review queue before a model build occurs, the queue initially serves up random documents because no scores are available for documents until a model build has completed.

3.4.2 Special considerations

- In order to make the Coverage Review even more efficient, we recommend suppressing duplicate documents from your Active Learning project on project setup. For more information, see Creating a new Active Learning project on page 14.
- The Active Learning model only builds when you have at least five documents coded with the positive choice and five coded with the negative choice.

3.4.3 Starting the Coverage Review queue

To start the Coverage Review queue:

1. Add reviewers to the queue.

   Note: We recommend no more than 150 concurrent reviewers per project. Concurrent reviewers are defined as reviewers making coding decisions in an Active Learning queue at the same time. There is no limit to how many reviewers you can add to a queue as long as the number of concurrent reviewers remains at 150 or fewer.

   a. Click **Add Reviewers** in the queue modal.

   b. Find and select an individual reviewer, or type to enter a reviewer name. You can also click **All Users** to select every user in the group.

c. Click the green check mark to save your changes.

d. Click the red X to cancel reviewer changes.

2. Click **Start Coverage Review**.

The next time an active reviewer accesses the view for the Active Learning project, they see a blue banner which indicates they have access to the queue. For more information, see Reviewer access on page 19.

You can see the following items in the Coverage Review queue modal:

- **Coded** - the number of documents coded and skipped in the Coverage Review queue.
- **Active Reviewers** - the number of reviewers added to the project.

Admins can pause the review at any point in the project by clicking the **Pause Coverage Review** button on the queue modal. Once the review is paused, the access point on the document view is disabled. You can restart the review at any time. Documents coded on the review field while the review is paused are included the next time the model rebuilds.

3.4.3.1 Adding Reviewers

From the project homepage, admins can add reviewers to each queue on a user-by-user basis or in bulk.

Note: After you ARM an Active Learning project, you must re-add reviewers to the queue.

To add reviewers:

1. Click **Add Reviewers** in the queue modal.
2. Find and select an individual reviewer, or type to enter a reviewer name. You can also click **All Users** to select every user in the group.
3. Click the green check mark to save your changes.
4. Click the red X to cancel reviewer changes.

Note: We recommend no more than 150 concurrent reviewers per project. Concurrent reviewers are defined as reviewers making coding decisions in an Active Learning queue at the same time. There is no limit to how many reviewers you can add to a queue as long as the number of concurrent reviewers remains at 150 or fewer.

3.4.4 Adding new documents

If you want to add or remove documents from your Active Learning project after review has begun, complete the following:

1. Click **Pause Review** to stop the review queue.
2. Add or remove documents to the saved search used as the data source for the classification index. Be sure to return only the extracted text in your saved search.
3. Navigate to the **Analytics indexes** tab, and then click the classification index used to create your Active Learning project.
4. On the index console, click **Run**, then **Incremental**.
5. Once the index finishes populating, return to your review queue and click **Start Review**.

Notes:

- All documents, including the newly added documents, are given a rank score after the incremental population.
- The following explains the way Active Learning handles documents removed from the project:
  - Documents in a queue that have been coded and then deleted from the workspace are marked as skipped. If the document was manually coded, then deleted, the manually coded statistics update to the correct values.
  - When the index is repopulated, deleted documents and documents removed from the search are removed from the model. After repopulation, the Project Size statistics and the count of coded documents are updated.
  - If Project Validation is in progress when an index is repopulated, the Project Validation is canceled.

4 Best practices for Active Learning review

This article is intended as a reference for reviewers in an Active Learning project. We've included best practices and overall considerations for working with computer-assisted review.

4.1 General coding tips

We recommend the following guidelines when reviewing documents:

- **Consistency is crucial** - consult fellow reviewers on difficult coding decisions to make sure your team is coding consistently.
- **Don't touch** - never add choices to the Designation tag. If you don't see the designation you need or you have a technical problem, ask the project administrator for a solution.
- **Double check** - always check the extracted text of a document to be sure it matches the content in other views. Whenever possible, review from the Extracted Text viewer.
- **When in doubt, ask** - if something confuses you, don't guess. Ask a system admin or project manager about the right course of action.

4.2 Coding according to the "four corners" rule

Active Learning predicts which documents will be responsive or nonresponsive based on the contents of the document itself. Family members, date range, custodian identity, and other outside factors do not affect the rankings. Because of this, the Active Learning engine learns best when documents are coded based only on text within the four corners of the document.

When you code documents in the Active Learning queue, you are both coding the document and teaching the engine what a responsive document looks like. Therefore, your coding decisions should follow the "four corners" rule: code only based on text within the body of the document, not based on surrounding factors.

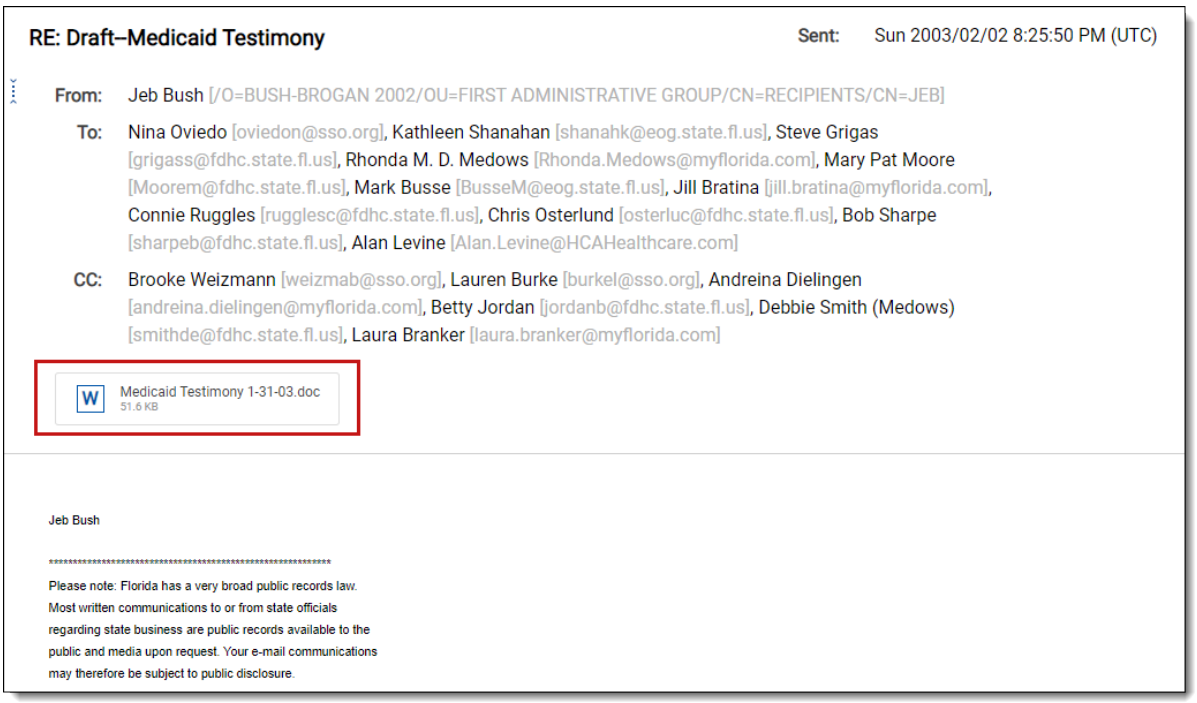
Having a one or two documents which fail this rule will not harm the overall accuracy of Active Learning's predictions. However, if you want to track large numbers of documents which are responsive for reasons outside of the four corners, we recommend talking to the project manager about setting up a secondary coding field for those. Because the secondary coding field is not used by Active Learning, it will not interfere with training the engine.

4.2.1 Common scenarios which fail the "four corners" rule

The following scenarios violate the "four corners" rule, and will **not** train the Active Learning engine well:

- The document is a family member of another document which is responsive.
- The document comes from a custodian whose documents are presumptively responsive.
- The document was created within a date range which is presumptively responsive.
- The document comes from a location or repository where documents are typically responsive.

For example, the following email has a responsive attachment. However, the body of the email—the text within the four corners—is only a signature and a privacy disclaimer. Because the body of this email is not responsive, this document does not pass the "four corners" rule.



RE: Draft--Medicaid Testimony

Sent: Sun 2003/02/02 8:25:50 PM (UTC)

From: Jeb Bush [/O=BUSH-BROGAN 2002/OU=FIRST ADMINISTRATIVE GROUP/CN=RECIPIENTS/CN=JEB]

To: Nina Oviedo [oviedon@sso.org], Kathleen Shanahan [shanahk@eog.state.fl.us], Steve Grigas [grigass@fdhc.state.fl.us], Rhonda M. D. Medows [Rhonda.Medows@myflorida.com], Mary Pat Moore [Moorem@fdhc.state.fl.us], Mark Busse [BusseM@eog.state.fl.us], Jill Bratina [jill.bratina@myflorida.com], Connie Ruggles [rugglesc@fdhc.state.fl.us], Chris Osterlund [osterluc@fdhc.state.fl.us], Bob Sharpe [sharpeb@fdhc.state.fl.us], Alan Levine [Alan.Levine@HCAHealthcare.com]

CC: Brooke Weizmann [weizmab@sso.org], Lauren Burke [burkel@sso.org], Andreina Dielingen [andreina.dielingen@myflorida.com], Betty Jordan [jordanb@fdhc.state.fl.us], Debbie Smith (Medows) [smithde@fdhc.state.fl.us], Laura Branker [laura.branker@myflorida.com]

Medicaid Testimony 1-31-03.doc
51.6 KB

Jeb Bush

****

Please note: Florida has a very broad public records law. Most written communications to or from state officials regarding state business are public records available to the public and media upon request. Your e-mail communications may therefore be subject to public disclosure.

4.3 Factors which affect Active Learning predictions

Not all responsive documents will inform the Active Learning engine equally. The following factors affect how much the engine will learn from each document:

- **Sufficient text** - if there are only a few words or short phrases in a document, the engine will not learn very much from it.
- **Email headers and signatures** - these are filtered out automatically by Active Learning. They are still visible, but they are not considered for Active Learning's predictions.
- **Images** - text contained only in images, such as a photograph of a contract, cannot be read by Active Learning. The system works only with the extracted text of a document.
- **Numbers** - numbers are not considered by the Active Learning engine.

5 Monitoring an Active Learning project

The actual Active Learning process takes place using a combination of time and detection of new coding decisions. After the Active Learning model completes its first build (once you have at least five documents coded with the positive choice and five coded with the negative choice), it builds at maximum every 20 minutes after the previous build to include coding decisions not included in the most recent build. If reviewer activity has been idle for five minutes and there are coding decisions not included in the most recent build, the model will start a build.

There are a number of ways to monitor the progress of an Active Learning project using tools on the Project Home and Review Statistics tabs.

Note: Each time an admin accesses the Project Home page - via a page refresh or from a different page - the latest data will reflect in the Project Home display.

See these related pages:

Review statistics on page 56

5.1 Project Home dashboard

The Project Home dashboard gives a high-level overview of the documents in your Active Learning project. After you first create the project, the dashboard displays the Project Size and coding statistics based on the pre-coded documents in your data source, if they exist.

Use the Project Home dashboard to understand the following:

- How many documents have been coded in your project.
- How many documents have been coded outside the queue (manually-selected documents).
- How many documents have been skipped.

The project home dashboard contains the following statistics:

- **Project Size** - the number of documents in the project. This count reflects the documents successfully indexed. Documents that were removed during the index build are excluded from this count.
- **Coded [Positive Choice]** - the number of documents coded with the positive designation on the review field. This count includes documents coded in the review queue and manually-selected documents. These documents are used to teach the model.
  - **Positive Choice manually-selected** - the number of documents that were coded on the positive choice designation field outside of the review queue.
- **Coded [Negative Choice]** - the number of documents coded with the negative designation in the review field. This count includes documents coded in the review queue and manually-selected documents. These documents are used to teach the model.
  - **Negative Choice manually-selected** - the number of documents that were coded on the negative choice designation field outside of the review queue.
- **Skipped** - the number of documents served to a reviewer that weren't coded by an end reviewer.

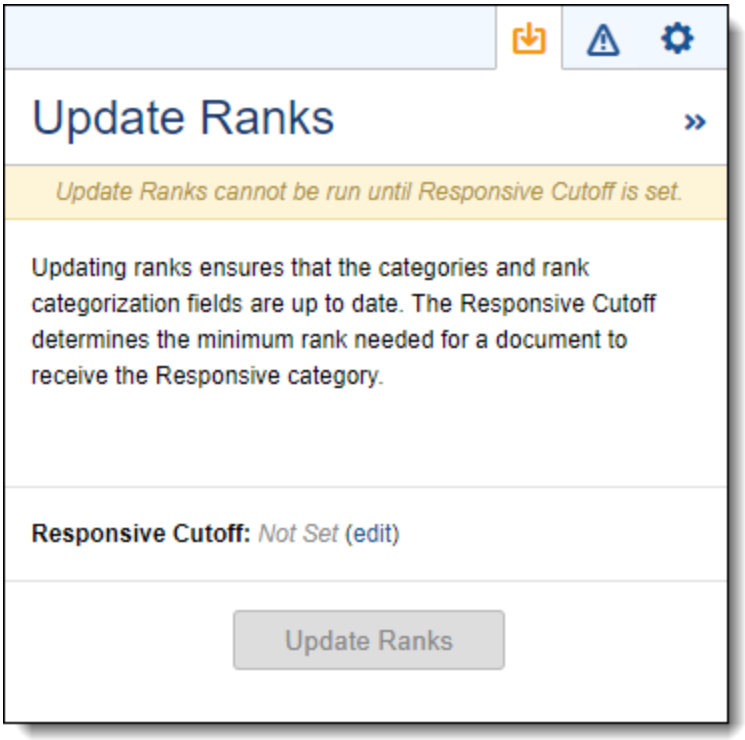
Note: If a coding decision is updated on a document reviewed in the queue, it will not change to a manually selected document.

5.2 Update ranks

In the top-right corner of the Project Home, you'll see three icons. The first icon is the Update Ranks icon . You can use Update Ranks to manually-update the document ranks and ensure the rank categorization field is up to date. Click **edit** to update the <Positive Choice> Cutoff value. Note that updating ranks doesn't change the coding decision on the document.

Note: When you update the <Positive Choice> Cutoff value, the value is updated in all three places where it's used in the application: Project Validation, Update Ranks, and Project Settings.

Active Learning ranks all documents from 100 (most likely to be in the positive category) to 0 (most likely to be in the negative category). These document ranks drive the Prioritized Review and Coverage Review queue logic, but they can also be imported into the Document object for other workflows. For instance, you could batch out to reviewers all documents above a certain rank. Because it is a SQL-intensive operation, ranks are only stored on the document object on demand. Since ranks change, at least slightly, with each model build, a project administrator can choose to update these ranks on demand.



In order to update ranks, you must specify the <Positive Choice> Cutoff. This value determines the minimum rank needed for a document to be predicted as responsive. You can look at the Document Rank Distribution chart to visualize how this will categorize your documents. Every document on or above the rank cutoff (to the right when looking at the Document Rank Distribution chart) receives the positive prediction, such as "responsive," while everything below the rank cutoff (to the left) receives the negative or

"non-responsive" prediction. Note that this predictive categorization is not the same as the coding decision on the document.

Note: When you update the <Positive Choice> Cutoff value, the value is updated in all three places where it's used in the application: Project Validation, Update Ranks, and Project Settings.

Clicking Update Ranks for the first time creates the following fields:

- **Categories - <Project Name> Cat. Set** - the positive or negative choice based on the <Positive Choice> Cutoff value. Documents below the cutoff are automatically given the negative choice category.
- **CSR - <Project Name> Cat. Set::Category Rank** - the rank score, on a scale from 0 to 100, from the last time update ranks was run.
- **CSR - <Project Name> Cat. Set::Category Name** - the positive or negative choice based on the <Positive Choice> Cutoff value. Documents below the cutoff are automatically given the negative choice category.

  Note: The values in the rank categorization fields are distinct from the coding decision on the document. For example, a document may be coded not responsive, but if the rank is at or above the <Positive Choice> Cutoff value, it is categorized as responsive.

The next time you update ranks, these fields are populated with the most up-to-date rank and category information. For more information on how to use these fields, see Quality checks and checking for conflicts on page 37.

While updating ranks, the Update Ranks icon displays a small clock icon in the corner. If you click the icon to open the fly-out modal, you can view the update progress. You can update ranks again only after the current update is complete.

5.2.1 Running a search on a classification index

If you don't want to run update ranks, you can instead run a search against a classification index to quickly return documents of a certain rank or within a range of ranks.

Notes:

- If you try running a search for the first time while the classification index is populating and building, the search will try to complete for five minutes. If the index hasn't finished building within this time, the search will fail. However, you can re-run the search after the index finishes.
- If you've previously run the search, these results are cached. If you try re-running the search while the classification index is building, you'll see these old, cached results. Once the index build completes, the results are refreshed with the latest index build results.

To run a search on a classification index:

1. Navigate to the **Documents** tab.
2. From the search bar, select the classification index associated with your Active Learning project.

   Note: The index you select must be associated with an Active Learning project that has been built (at least five documents coded with the positive designation and five coded with the negative designation).

3. Using the next drop-down, select whether to search for **Greater than or equal to**, **Less than or equal to**, **Between**, or **Is** the rank value you enter.
4. Click **Search**.

The Rank column displays rank results relevant to your search for the most recent index build. This differs from the **CSR- <Project Name> Cat. Set::Category Rank** field generated by Update Ranks, which stores old results until you manually re-run.

The rank scores are rounded to two decimal places. Note that these results are temporary, and you can't run the mass operations Sum, Tally, and Average on them.

5.3 Document rank distribution

The Document Rank Distribution chart shows the number of documents at each rank, from 0 to 100, grouped by the reviewers' coding decisions on those documents. A relevance rank near zero indicates the model believes the document is more likely coded on the negative review field choice. On the other hand, a rank closer to 100 means the model believes a document is more likely to be coded on the positive review field choice.

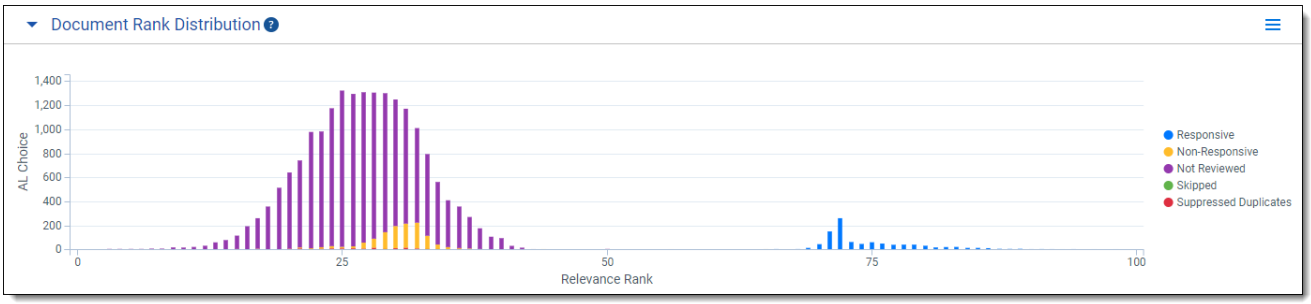
Note: If you see a disproportionately high number of documents ranked at 50, this may indicate that you have a high number of empty documents in your project. For more information, see Troubleshooting a large group of documents ranked at 50.

Use the rank distribution chart to understand the following:

- The number of predicted relevant documents that remain for review
- The agreement between reviewers and the Active Learning model
- The number of documents the model does not understand well

The charts reports documents reviewed from the Prioritized Review or Coverage Review queue, as well as documents coded outside of the queue. In the early stages of an Active Learning project, most documents will have a relevance rank score between 40 and 60 until the model begins training. As the model learns throughout the project life cycle, the Rank Distribution chart is expected to gravitate toward 0 or 100 depending on how documents are coded on the positive choice or negative choice. Typically, a well trained model will have a large pile of documents centered around 70 or 75, and another larger pile centered around 20 or 25. Documents with scores in the 70s are very likely relevant, and documents in the 20s are very likely not relevant.

The review state of the documents are also overlaid on this distribution. Note that it is possible for a document coded on the positive choice to have a lower relevance ranking; this is because the rank is simply the model's prediction.



You can view the following colors on the chart:

- **Blue (Positive Choice)** - documents coded on the positive choice review field.
- **Yellow (Negative Choice)** - documents coded on the negative choice review field.
- **Purple (Not Reviewed)** - documents within the project's scope, but not yet coded. Their rank is based on the Active Learning model's predictions.
- **Green (Skipped)** - documents that were skipped.
- **Red (Suppressed Duplicate)** - documents that are suppressed by the suppress duplicates setting after a textually similar document has been coded.

Note: When a full population is performed, all previously identified suppressed documents are marked as "Not Reviewed" in the Document Rank Distribution chart.

You can interact with the Document Rank Distribution Chart to hide the different categories of documents to more easily view the particular categories of documents that remain in the chart. For example, to hide the Not Reviewed documents, click the Not Reviewed label on the legend. Upon clicking, the bar chart will rescale for the remaining documents.

Active Learning: Large group of documents ranked at 50

If you see a disproportionately high number of documents ranked at 50, this may indicate that you have a high number of empty documents in your project. This could be because you indexed the wrong field, or your project includes a lot of images or media files, which don't contain words. Documents without any words can't be ranked by the model, so it ranks them all at 50.

Note: Documents that consist only of words that are not present in coded documents will also not be rankable by the model.

Use the following methods to try to identify documents without text:

- Ensure the saved search you used to create the classification index returns only the extracted text field. If it doesn't, update the index and click **Run**, then **Full** to replace the previous results. Once the index is rebuilt, the model rebuilds and all documents are reclassified.
- Search for documents with zero-length text by using a field like **Extracted text size**. Remove these documents from your saved search by including a condition:
  - **Field:** Extracted text size
  - **Operator:** is greater than
  - **Text box:** 0

  Click **Run**, then **Full** to replace the previous results.

If neither of these methods work, run **Update ranks** if you haven't done so recently. Then, search through the documents with rank score of 50 for other patterns. If you find legitimate reasons why documents would be empty (ex. image files whose extracted text field contains only a URL), then systematically exclude them from the saved search and incrementally populate your index.

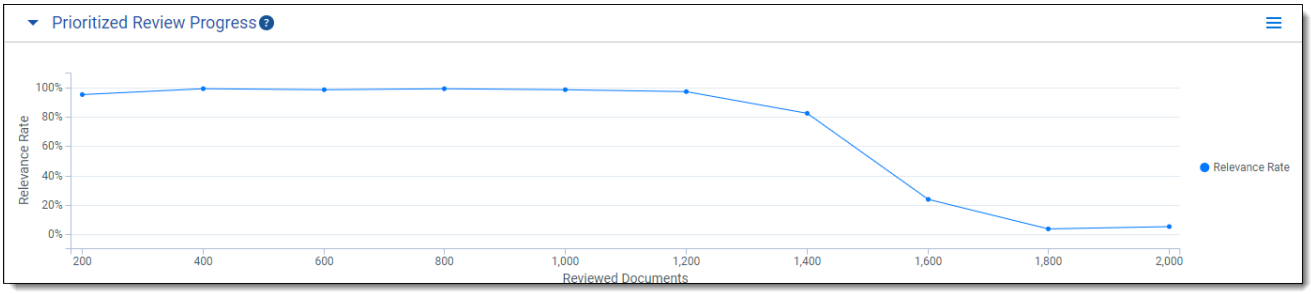
5.4 Prioritized review progress

The Prioritized Review Progress chart tracks the model's ability to find relevant documents over time. It does this by monitoring reviewers' coding decisions on the high-ranking documents chosen by the queue. Because the Prioritized Review queue focuses on high-ranking, uncoded documents, over time it will have to serve up lower and lower ranks, which corresponds to less relevant documents.

As a general trend, you should expect to see the relevance rate decline over the course of the review. However, you may see a spike in the relevance rate if a large amount of new documents are added to the project, or if the definition of relevance changes during the course of the review.

Notes:

- This chart only updates when documents are coded within the Prioritized Review queue. If only manually-selected documents are coded, the chart won't display relevance rate data.
- The relevance rate calculation does not include family documents or documents included in the queue for index health.
- This measurement is not cumulative with regard to the entire document set.



The x-axis charts documents in groups of 200 which have been reviewed in the Prioritized Review queue. Data won't appear on the chart until at least 200 documents are coded. Once 200 documents are coded in Prioritized Review, relevance rate data appears on the chart. A new data point appears every time another 200 documents are coded.

The y-axis charts the relevance rate. This is the percentage of the highly ranked documents which are confirmed by the reviewer to be relevant.

5.5 Quality checks and checking for conflicts

We recommend running ongoing quality checks over the course of the project. The Active Learning process is fairly tolerant of some inaccurate or inconsistent coding, but it's good practice to monitor your queue for conflicts between reviewers and the Active Learning model.

5.5.1 Useful dashboards

You can make custom widgets and dashboards with various active learning fields. Some of the fields are created when you create your Active Learning project, while others are created the first time you update ranks. A few fields in particular may be of use to you:

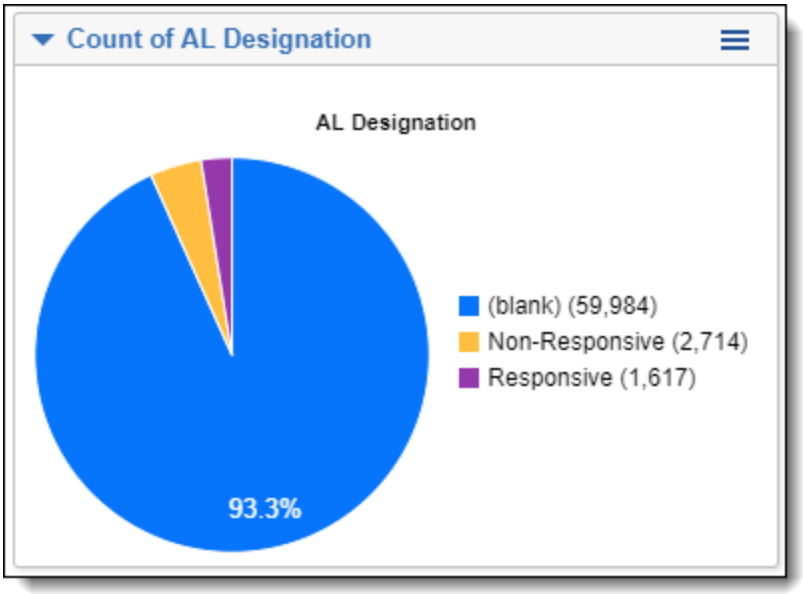
- **CSR - <Project Name> Cat. Set::Category Rank** - the rank score, on a scale from 0 to 100, from the last time update ranks was run.

- **Categories - <Project Name> Cat. Set** - the positive or negative choice based on the <Positive Choice> Cutoff value from the last time update ranks was run. Documents below the cutoff are automatically given the negative choice category.
- **<Project Name> Reviewers::User** - the name of the reviewer who coded the document in the queue.
- **<Project Name>:: Prioritized Review** - the 200-document interval in which the document was reviewed in the Prioritized Review.
- **<Project Name>:: Coverage Review** - the 200-document interval in which the document was reviewed in the Coverage Review.
- **<Project Name>:: Project Validation** - the Project Validation in which the document was reviewed.

5.5.1.1 Review field coding

This widget tracks documents manually coded by reviewers.

Note: The AL Designation field could be named Responsive or Relevance but for this example we have created a single choice field named AL Designation.



To create this widget, complete the following:

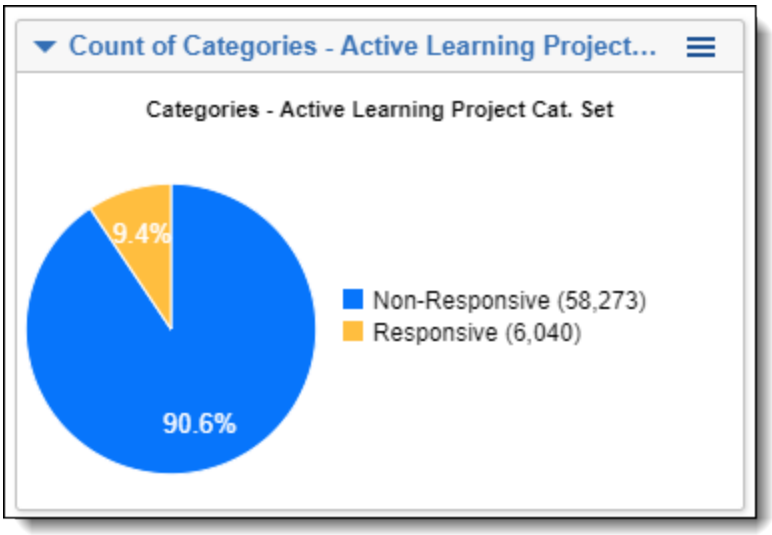
1. From the **Documents** tab, click **Add Widget**.
2. Select **Pivot**.
3. Complete the following fields:
   - **Group by**: AL Designation
   - **Group By Results Returned**: All
   - **Pivot on**: <Grand Total>

- **Sort on**: <Grand Total> (Descending)
- **Default Display Type**: Pie Chart

4. Click **Add Pivot**.
5. After adding the pivot, click the icon at the top right of the Pivot widget to display the Pivot options pane.
6. Enable the following options:
   - Show (blank)
   - Show Legend

5.5.1.2 Active Learning Results Summary

This widget will help you understand the system identified responsiveness.



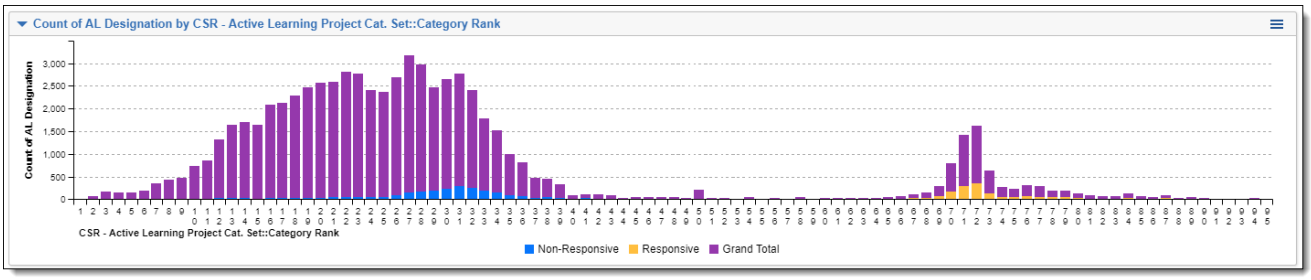
To create this widget, complete the following:

1. From the **Documents** tab, click **Add Widget**.
2. Select **Pivot**.
3. Complete the following fields:
   - **Group by**: Categories - Active Learning Project Cat. Set
   - **Group By Results Returned**: All
   - **Pivot on**: <Grand Total>
   - **Sort on**: <Grand Total> (Descending)
   - **Default Display Type**: Pie Chart
4. Click **Add Pivot**.

5. After adding the pivot, click the icon at the top right of the Pivot widget to display the Pivot options pane.
6. Enable the following options:
   - Show Legend

5.5.1.3 Machine Classification against Coding Values

This widget compares the human coding values against Active Learning classification.



To create this widget, complete the following:

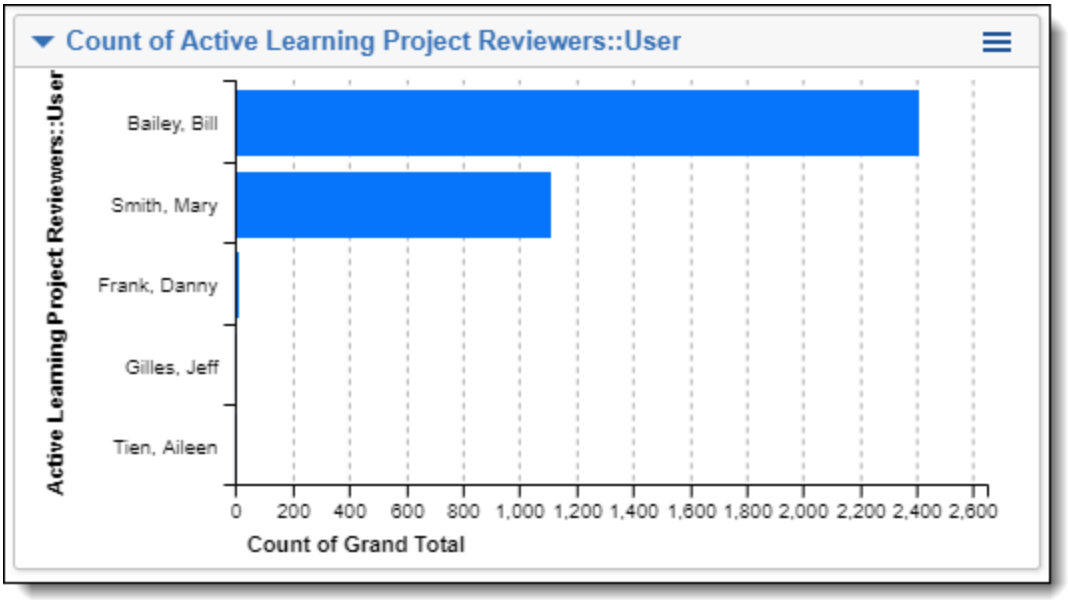
1. From the **Documents** tab, click **Add Widget**.
2. Select **Pivot**.
3. Complete the following fields:
   - **Group by**: CSR - Active Learning Project Cat. Set::Category Rank
   - **Group By Results Returned**: All
   - **Pivot on**: AL Designation
   - **Pivot on Results Returned**: All
   - **Sort on**: CSR - Active Learning Project Cat. Set::Category Rank (Ascending)
   - **Default Display Type**: Stacked Bar Chart
4. Click **Add Pivot**.
5. After adding the pivot, click the icon at the top right of the Pivot widget to display the Pivot options pane.
6. Enable the following options:
   - Show (blank)
   - Show Legend
   - Vertical Orientation

Notes:
This chart looks similar to the document rank distribution chart on the Active Learning project homepage, with the following differences:

- The colors for responsives and non-responsives may be different from the document rank distribution chart.
- Any suppressed duplicates appear together with the rest of the uncoded documents.
- The ranks do not update automatically. They will only refresh when you run Update Ranks from the project homepage.

5.5.1.4 User Coding Tracker

This widget will show the number of documents coded per reviewer. This is handy to understand who is coding what and track progress.



To create this widget, complete the following:

1. From the **Documents** tab, click **Add Widget**.
2. Select **Pivot**.
3. Complete the following fields:
   - **Group by**: Active Learning Project Reviewers::User
   - **Group By Results Returned**: All
   - **Pivot on**: <Grand Total>
   - **Sort on**: <Grand Total> (Descending)
   - **Default Display Type**: Bar Chart
4. Click **Add Pivot**.

5. After adding the pivot, click the icon at the top right of the Pivot widget to display the Pivot options pane.
6. Enable the following options:
   - Show Legend
   - Horizontal Orientation

5.5.2 Using these widgets

After creating these widgets, use the following workflows:

- Check for conflicts by doing the following:
  - Select the Responsive items coded by reviewers. Select the Non Responsive classified items. Consult the Reviewers::User widget to see who coded these documents if they were coded in a queue.

    Note: These are documents that Active Learning is predicting as non-responsive, but the reviewer has coded as responsive. While such conflicts usually cannot be eliminated entirely, too many of them could be a sign of reviewer inconsistency, not coding on the "four corners" of the document, or a saved search for the index which returns something other than the extracted text field.

  - Select the Non-Responsive items coded by reviewers. Select the Responsive classified items. Consult the Reviewers::User widget to see who coded these documents if they were coded in a queue.

    Note: These are documents that Active Learning is predicting as responsive, but the reviewer has coded as non-responsive. While such conflicts usually cannot be eliminated entirely, too many of them could be a sign of reviewer inconsistency, not coding on the "four corners" of the document, or a saved search for the index which returns something other than the extracted text field.

  - Run a search for documents **Greater than or equal to** your rank cutoff. Then, filter on the **Designation** field to return documents within this rank that were coded on the negative designation.
- Select a reviewer. See how the coding values change to understand how the coding breakdown occurs per person. See if someone is coding a much higher or lower percentage compared to other users.

5.6 Determining when a project is done

When the project is nearing completion, it can become more difficult to find additional relevant documents. As you monitor your project, here are some tips for determining when a project is done:

- **Document rank distribution** - look for few or no purple bars (Not Reviewed) over the 50 mark for relevance.
- **Prioritized review progress** - you should see a decline in the prioritized review progress line chart indicating that very few responsive documents are being found.

- **Richness** - if you estimated richness at the start of your project, you have a rough estimate of how many relevant documents are in your project. The number of documents coded relevant should be close to the richness estimate when the project is done.

Once you believe you are near the end of your project, you can run Project Validation to confirm the state of the project. For more information, see Project Validation and Elusion Testing on the next page.

6 Project Validation and Elusion Testing

Please note: For the initial Server 2022 release, the only available project validation type was **Elusion Only**. The **Elusion with Recall** validation type, which provides richness, recall, and precision statistics, is now available as part of the Server 2022 installation files.

- To install the **Elusion with Recall** feature, visit the Community site for the Relativity Server 2022 GA installation files and follow the instructions for Relativity Precision, Richness, and Recall in Active Learning.
- For instructions specific to the **Elusion Only** workflow, see Elusion Test on the Server 2021 documentation site.

Project Validation is used to validate the accuracy of an Active Learning project. The goal of Project Validation is to estimate the accuracy and completeness of your relevant document set if you were to stop the project immediately and not produce any unreviewed documents below the rank cutoff. The primary statistic, elusion rate, estimates how many low-ranked uncoded documents are actually relevant documents that you would leave behind if you stopped the project. The other validation statistics give further information about the state of the project. We recommend running Project Validation when you believe the project has stabilized and the low-ranking documents have an acceptably low relevance rate. However, you can run Project Validation at any point after the first model build.

When you run Project Validation, you must first specify the rank cutoff. Then you specify the confidence level for elusion, and either the number of documents in the sample or the margin of error for elusion. Based on these inputs, a sample will be taken. If you choose Elusion Only, the elusion sample is taken from the uncoded documents below the specified rank cutoff. The uncoded documents include documents that were skipped, suppressed, or never reviewed. If you choose Elusion with Recall, the sample is taken from all documents in the Active Learning project, regardless of rank or coding status. Reviewers then code the uncoded sample documents on the project review field, and the results are used to calculate project validation statistics.

For a detailed explanation of how the validation statistics are calculated, see Project Validation statistics on page 49.

Note: Project Validation does not check for human error. We recommend that you conduct your own quality checks to make sure reviewers are coding consistently. For more information, see Quality checks and checking for conflicts on page 37.

6.1 Key definitions for Project Validation

The following definitions are useful for understanding Project Validation:

- **Discard pile** – the set of uncoded documents with ranks below the rank cutoff.
- **Discard-pile elusion rate** – the percentage of documents in the discard pile that are relevant. It's not possible to calculate this number precisely (with zero error) without coding every document in the discard pile. Therefore, we use sampling to estimate the discard-pile elusion rate. This gives us a sample elusion rate along with a margin of error and confidence level, which capture the amount of uncertainty in the estimate.
- **Sample elusion rate** – the percentage of documents in the elusion sample that are relevant.

- **Pre-coded documents in sample** - documents which are included in the sample but not served up for coding. When Elusion with Recall is selected, it samples across the entire Active Learning Project, including from documents that already have coding decisions. These pre-coded documents are assumed to be correctly coded, and they are not served up to reviewers for re-review in the validation queue.
- **Rank Cutoff** - the numeric cutoff for positive prediction. This is fixed before Project Validation begins.

6.2 Choosing a validation type

The key differences of each validation type are listed below. If you aren't sure which type to choose, we recommend Elusion with Recall.

Elusion with Recall - samples all documents, regardless of rank or coding status, and calculates elusion rate, richness, recall, and precision. This has the following advantages:

- The richness estimate helps you interpret the elusion rate. For more information, see Project Validation statistics on page 49.
- The recall estimate is often requested by receiving parties, especially government regulators.
- If you run validation near the end of a prioritized review, this validation type takes roughly the same effort as Elusion Only, but gives a clearer picture of the project state. Because most high-ranking documents in the sample have already been coded, the reviewers will be coding a similar number of documents to an Elusion Only test.

Elusion Only - samples uncoded documents below the rank cutoff and calculates only the elusion rate. Because it samples from only the discard pile, Elusion Only requires less coding from reviewers when you have many uncoded documents above the cutoff. If you run validation early in your project, this validation type will be faster, but it gives less information.

Note: Documents which were previously coded will be sampled for Elusion with Recall's statistics, but they will not be served up to reviewers a second time. They are not sampled at all for Elusion Only tests because they are already coded, and therefore did not elude review.

6.3 Starting Project Validation

The Project Validation queue appears along with the other review queues after a new project is created. Starting Project Validation disables all other active queues in the project and suspends model updates until Project Validation is completed.

To run Project Validation, complete the following:

1. Click **Add Reviewers** on the Project Validation card and confirm you want to start Project Validation.

   Note: It is common practice to restrict project validation to a small set of knowledgeable reviewers. This typically includes fewer than ten reviewers.

2. Wait for the system to set up the test. Once the card reads **Start Project Validation**, click the queue.
3. On the Project Validation setup window, complete the following fields:
   - **<Positive Choice> Cutoff** - the rank which divides positive and negative predictions, such as "relevant" and "not relevant." This cutoff is used throughout Active Learning and factors heavily into the validation statistics.

Note: When you update the rank cutoff value, the value is updated in all three places where it's used in the application: Project Validation, Update Ranks, and Project Settings.

- **Validation Type**
  - **Elusion Only** - starts validation which checks the elusion rate only.
  - **Elusion with Recall** - starts validation which checks elusion rate, richness, recall, and precision.
- **Sample Type**
  - **Statistical** - creates a random sample set of a size that is based on a given Confidence and Margin of Error for elusion.
  - **Fixed** - creates a random sample of a fixed number of documents. This is a sample of the discard pile for Elusion Only and a sample of the entire project for Elusion with Recall.
- **Elusion Confidence Level** - the probability that the sample elusion rate is a good estimate of the discard pile elusion rate (i.e., within the margin of error). Selecting a higher confidence level requires a larger sample size.
- **Est. Elusion Margin of Error** - the maximum difference between the sample elusion rate and the discard-pile elusion rate. Selecting a lower margin of error requires a larger sample size.

Note: Elusion Only tests keep the margin of error that was chosen at the beginning of the test. For Elusion with Recall tests, the actual margin of error is recalculated at the end of project validation. It may be narrower than the original estimate.

- **Reviewers** - the users that will review documents during Project Validation.

4. Click the green check mark to accept your settings.
5. Click **Start Review**.

6.4 Tracking sampled documents

If you want to run your own calculations on the Project Validation sample, you can create lists to track the sampled documents. This process is optional.

Follow the steps below to create three lists in the Documents view:

1. **Pre-coded documents**
   1. After you start Project Validation, but before reviewers have coded any documents, search for documents which already have the <project name>::Project Validation field set. These are the pre-coded documents.
   2. Save the results as a list called "Validation Pre-coded" or similar.
2. **High-ranking Project Validation queue documents**
   1. After coding is complete, search for documents which have the <project name>::Project Validation field set, but which do not belong to the "Validation Pre-coded" list.

2. Add a search condition narrowing the results to documents which have a CSR - <project name> Cat. Set::Category Rank value greater than or equal to your cutoff value.
3. Save the results as a list called "Validation Predicted Positive" or similar.

3. **Low-ranking Project Validation queue documents**

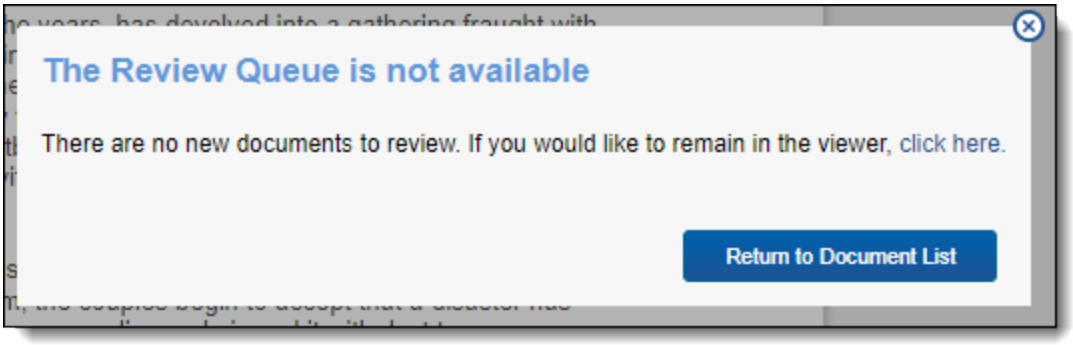
1. After coding is complete, search for documents which have the <project name>::Project Validation field set, but which do not belong to the "Validation Pre-coded" list.
2. Add a search condition narrowing the results to documents which have a CSR - <project name> Cat. Set::Category Rank value lower than your cutoff value.
3. Save the results as a list called "Validation Predicted Negative" or similar.

For more information on creating lists, see Mass Save as List in the Admin Guide.

6.5 Running Project Validation

Project Validation statistics are reported in Review Statistics and are updated during Project Validation. You can cancel Project Validation at any time. You can also pause a review queue by clicking the **Pause Review** button.

Reviewers access the queue from the document list like all other queues. Reviewers code documents from the sample until all documents have been served, at which point the following message appears:



Note: For best results, we strongly recommend coding every document in the Project Validation queue and avoiding skipping documents. For more information, see How Project Validation handles skipped documents on page 55.

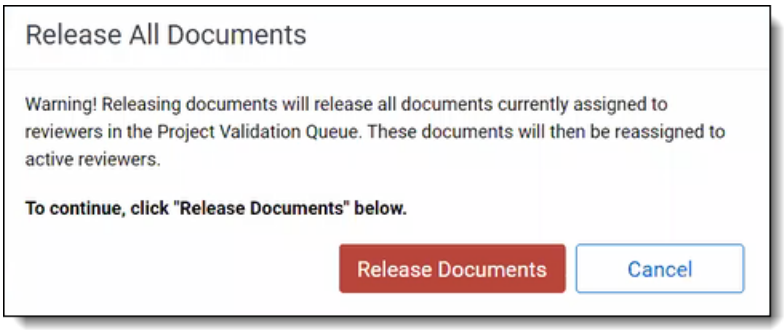
When a reviewer saves a document in Project Validation, the document is tagged in the **<Project Name>:: Project Validation** multi-choice field.

6.5.1 Releasing unreviewed documents

If a reviewer falls inactive and does not review the last few documents in a Project Validation queue, you can release those documents to be checked out by another reviewer. This option appears when there are four or fewer documents remaining in Project Validation.

To release unreviewed documents, select the following:

1. In the Active Learning Project Home, click **Release All Documents** on the Project Validation card.
2. Click **Release Documents** on the confirmation modal. The documents are now free to be checked out by active reviewers.



Note: The Release All Documents link remains on the Project Validation card after the documents have been released.

6.6 Reviewing Project Validation results

Once reviewers code all documents in the sample, you can access a summary of the Project Validation results by clicking **View Project Validation Results**. To view full results, click on the Review Statistics tab in the upper left corner of the screen. For more information, see Review statistics on page 56.

Based on the coding of the Project Validation sample, the results pop-up displays the following:

- **Elusion Rate** - the percentage of documents coded relevant in the uncoded, low-ranking portion of the sample. The elusion rate results are displayed as a range that applies the margin of error to the sample elusion rate, which is an estimate of the discard pile elusion rate. The rate is rounded to the nearest hundredth of a percent.

  Note: Documents that are skipped during the Project Validation queue are treated as relevant documents when calculating Elusion Rate. Therefore, coding all of the documents in the elusion sample guarantees the statistical validity of the calculated elusion rate as an estimate of the entire discard-pile elusion rate.

- **Eluded Documents** - the estimated number of eluded documents, calculated by multiplying the sample elusion rate by the number of documents in the discard pile. This number is subject to the final confidence and margin of error which can be found in Review Statistics.
- **Richness** - the percentage of relevant documents across the whole sample. This is calculated by dividing the number of positive-coded documents in the sample by the total number of documents in the sample. This allows us to predict a richness range for the whole project.
- **Recall** - the percentage of truly positive documents which were found by the Active Learning process. A document has been "found" if it was previously coded positive, or if it is uncoded with a rank at or above the cutoff.
- **Precision** - the percentage of found documents which are truly positive. A document has been “found” if it was previously coded positive, or if it is uncoded with a rank at or above the cutoff. Documents which were predicted positive but coded negative during validation will count against precision.

If documents were skipped during Project Validation, a warning appears on the modal. You can review these skipped documents, and they'll be reflected in the results as if they were coded during the test.

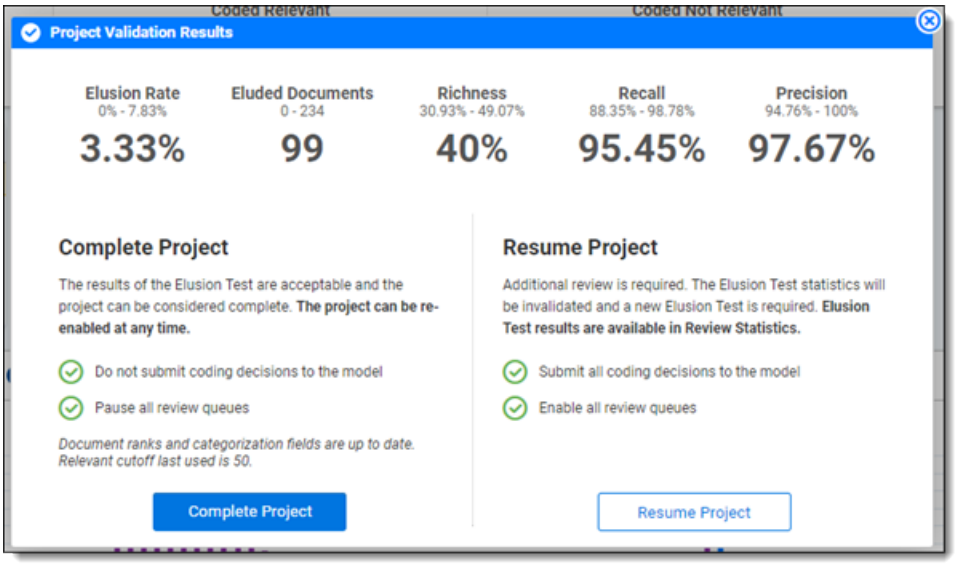
If you find the Project Validation results acceptable, select whether to **Update ranks upon completion**, and then click **Complete Project** to close the project. Once the project is complete, the model remains frozen. Any coding decisions that occur after Project Validation will not be used to train the (now frozen) model.

Note: Updating ranks upon accepting Project Validation results will use the Project Validation Rank Cutoff.

If you don't find the Project Validation results acceptable, click **Resume Project**, then click again to re-open the Active Learning project. This unlocks the model and allows it to rebuild. Any documents coded since Project Validation began, including those from the Project Validation queue itself, are included in the model build.

Full Project Validation statistics are reported in Review Statistics and persist after Project Validation is finished. This data is located under the Project Validation report tab.

See Review Statistics for more information.



6.7 Project Validation statistics

Please note: For the initial Server 2022 release, the only available project validation type was **Elusion Only**. The **Elusion with Recall** validation type, which provides richness, recall, and precision statistics, is now available as part of the Server 2022 installation files.

- To install the **Elusion with Recall** feature, visit the Community site for the Relativity Server 2022 GA installation files and follow the instructions for Relativity Precision, Richness, and Recall in Active Learning.
- For instructions specific to the **Elusion Only** workflow, see Elusion Test on the Server 2021 documentation site.

Active Learning's Project Validation feature provides four metrics for evaluating your review coverage. Together, these metrics can help you determine the state of your Active Learning project. Once you have

insight into the accuracy and completeness of your relevant document set, you can make an educated decision about whether to stop the Active Learning workflow or continue review.

For instructions on how to run Project Validation, see Project Validation and Elusion Testing on page 44.

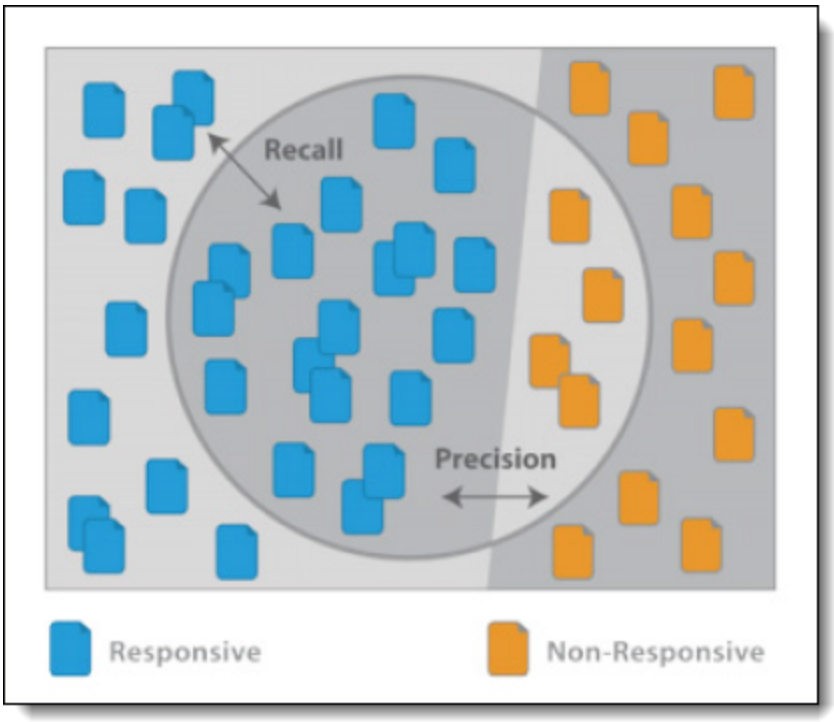
See these related pages:

Project Validation Strategies in Active Learning webinar

6.7.1 Defining elusion, richness, recall, and precision

Project Validation centers around four statistics, which are defined as follows:

- **Elusion rate** - the percentage of documents coded relevant in the uncoded, low-ranking portion of the sample. The elusion rate results are displayed as a range that applies the margin of error to the sample elusion rate, which is an estimate of the discard pile elusion rate. The rate is rounded to the nearest hundredth of a percent.
- **Richness** - the percentage of relevant documents across the whole sample. This is calculated by dividing the number of positive-coded documents in the sample by the total number of documents in the sample. This allows us to predict a richness range for the whole project.
- **Recall** - the percentage of truly positive documents which were found by the Active Learning process. A document has been "found" if it was previously coded positive, or if it is uncoded with a rank at or above the cutoff. Documents that were predicted negative but coded positive during validation will count against recall.
- **Precision** - the percentage of found documents which are truly positive. A document has been “found” if it was previously coded positive, or if it is uncoded with a rank at or above the cutoff. Documents that were predicted positive but coded negative during validation will count against precision.



Project Validation arrives at these values through its determination of true and false positives, as well as false negatives:

- **True Positive** - predicted relevant by the system and confirmed relevant during validation
- **False Positive** - predicted relevant by the system but coded not relevant during validation
- **False Negative** - predicted not relevant by the system but coded relevant during validation

For each of these metrics, Project Validation assumes that you will trust the human coding decisions over machine predictions, and that the prior coding decisions are correct. It does not second-guess human decisions.

Note: Project Validation does not check for human error. We recommend that you conduct your own quality checks to make sure reviewers are coding consistently. For more information, see Quality checks and checking for conflicts on page 37.

6.7.2 How Project Validation metrics are calculated

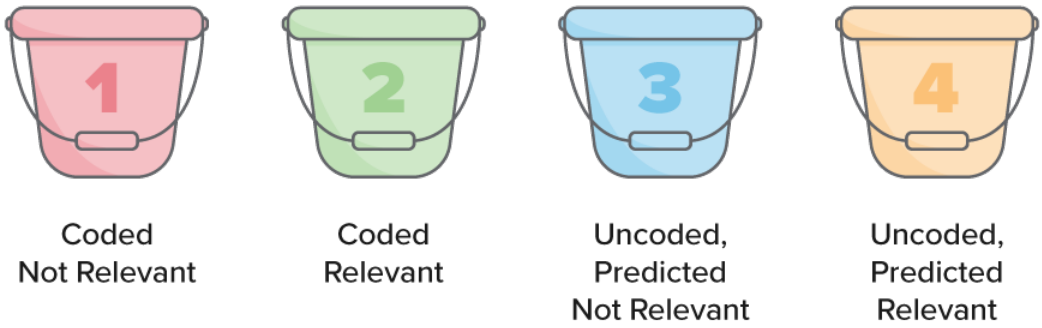
The Project Validation metrics center around two types of information:

- **Errors (false negatives and false positives)** - these are measured by elusion, recall, and precision.
- **Percentage of relevant documents** - this is measured by richness.

Each metric frames its information somewhat differently, depending on the numerator and denominator. In order to understand these differences, we can visualize them using the document groups defined below.

6.7.2.1 Document groups used for calculations

All documents in an Active Learning project can be categorized into the following four groups:



1. Documents which are coded not relevant
2. Documents which are coded relevant
3. Documents which are not coded, and predicted not relevant
4. Documents which are not coded, and predicted relevant

For the purposes of this explanation, we will refer to these groups as buckets.

A Project Validation sample contains documents from each bucket, and the proportions of the sample will roughly match the proportions of the population. For example, if 50% of your population falls into bucket 3, then the sample will also have roughly 50% of its documents drawn from bucket 3. In rare cases where the sample size is extremely small, such as 10 documents, or if almost no documents have been coded, then not all buckets may be represented.

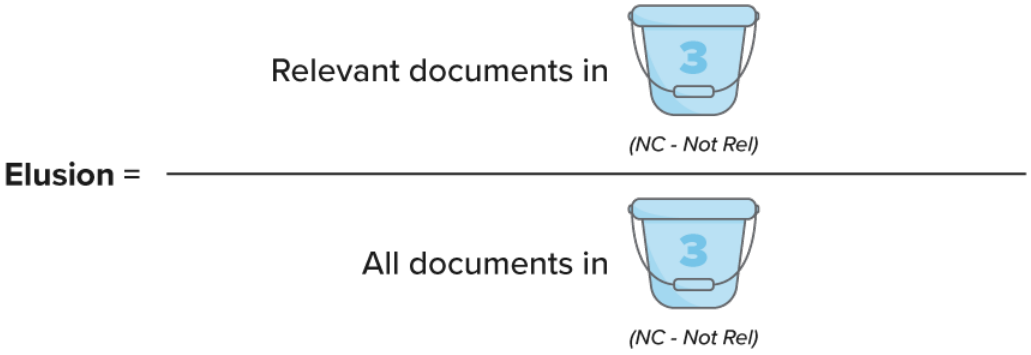
Note: These buckets are determined by a document's status at the start of Project Validation. For the purpose of these calculations, documents do not "switch buckets" during the course of validation.

6.7.2.2 Calculating the metrics based on document group

Using samples of the "buckets" of documents defined above, each of the Project Validation metrics can be calculated as follows:

- Elusion: Samples bucket 3
- Richness: Samples buckets 1, 2, 3, and 4
- Recall: Samples buckets 2, 3, and 4
- Precision: Samples buckets 2 and 4

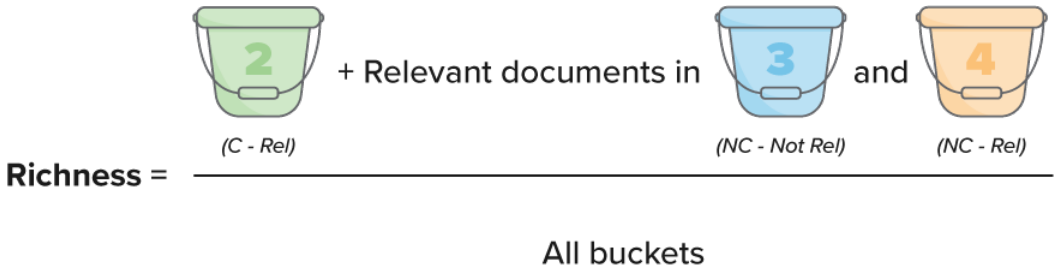
Elusion rate



Elusion = (Relevant documents in bucket 3) / (All documents in bucket 3)

Elusion measures the "error rate of the discard pile" — meaning, the relevant document rate in bucket 3. Documents which were coded relevant before starting project validation are not included in the calculation, regardless of their rank score.

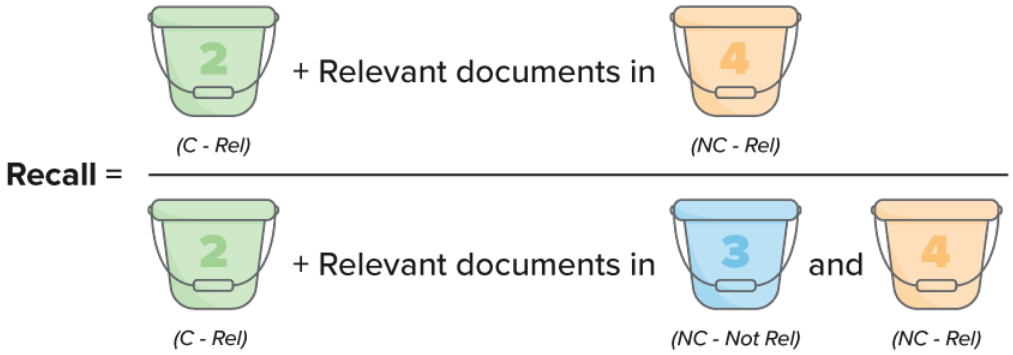
Richness



Richness = (Bucket 2 + any relevant documents found in buckets 3 and 4) / (All buckets)

Richness measures the overall relevance rate of all documents in the project. As an Active Learning project progresses, these relevant documents will be found in different buckets, but the overall percentage of relevant documents remains the same.

Recall

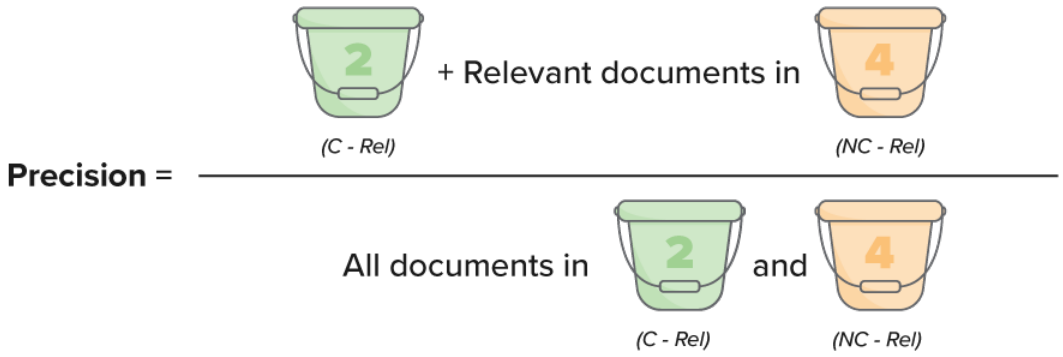


Recall = (Bucket 2 + relevant documents in bucket 4) / (Bucket 2 + relevant documents in buckets 3 and 4)

Recall measures the percentage of truly positive documents which were found by the Active Learning process. It is one of the most widely used metrics, but it is also the most complicated of the four to calculate. Unlike the other three metrics, it does not use complete buckets in the denominator, which makes it harder to visualize.

Recall shares a numerator with the precision metric, but the denominators are different. In recall, the denominator is "what is truly relevant;" in precision, the denominator is "what we are producing."

Precision



Precision = (Bucket 2 + relevant documents in bucket 4) / (All documents in buckets 2 and 4)

Precision measures the accuracy of what you're planning to produce. Unlike recall, it does not consider the low-ranking documents from bucket 3 that would be left out of a production.

Because bucket 2 grows larger and bucket 4 grows smaller as documents are coded, the precision rate will grow higher throughout the life of the project. Whenever you use the precision metric, remember that it considers all coded documents to be coded correctly. It does not account for possible human error.

6.7.3 How the Project Validation queue works

Once you start Project Validation, the system puts all sampled documents from buckets 3 and 4 into the queue for reviewers to code. This is the key to measuring Active Learning's predictions against human decisions. Documents which are coded during project validation do not switch buckets during the validation process; documents which started in buckets 3 and 4 are still considered part of 3 and 4 until validation is complete. This allows the system to keep track of correct or incorrect predictions when calculating metrics, instead of lumping all coded documents in with those which were previously coded.

For an explanation of the document "buckets," see Document groups used for calculations.

6.7.4 How the Project Validation sample is selected

Because all four metrics have different denominators, they all have different confidence intervals. Instead of selecting every single confidence interval and margin of error for your sample, Active Learning offers two simplified options which center around the elusion metric.

For an explanation of the document "buckets," see Document groups used for calculations.

6.7.4.1 Option 1: Fixed

When you choose the fixed sample size, you enter the total number of documents you want drawn for the sample. The setup window will calculate what percentage of that is being drawn from bucket 3, then estimate your elusion margin of error based on that. This estimated margin of error may change slightly once the true sample elusion rate has been measured during validation.

6.7.4.2 Option 2: Statistical

When you choose the statistical sample size, you enter the confidence level and margin of error you want for the elusion metric. The setup window will calculate how many sample documents to draw from bucket 3

for the elusion metric. If you chose Elusion with Recall, it will then calculate how many additional documents are needed to keep the other sampled buckets in proportion with bucket 3.

6.7.5 How Project Validation handles skipped documents

We strongly recommend coding every document in the Project Validation queue. Skipping documents lowers the randomness of the random sampling, which introduces bias into the validation statistics. To counter this, Active Learning gives conservative estimates. Each validation statistic counts a skipped document as an unwanted result.

The following table shows how skipped documents negatively affect each statistic.

Skipped Document	Effect on Elusion	Effect on Recall	Effect on Richness	Effect on Precision
Low-ranking	Increases elusion rate (Counts as relevant)	Lowers recall rate (Counts as non-relevant)	Raises richness estimate (Counts as relevant)	(No effect)
High-ranking	(No effect)	Lowers recall rate slightly (Counts as if it weren't present)	Raises richness estimate (Counts as relevant)	Lowers precision rate (Counts as non-relevant)

6.7.5.1 Documents skipped during Prioritized or Coverage Review

If a document was skipped during Prioritized Review or Coverage Review and is then served and coded during Project Validation, it will be designated as a coded document rather than a skipped document in that queue's Review Statistics tab.

7 Review statistics

The Review Statistics tab contains project statistics for each queue as well as a model update history.

7.1 Review Summary

The Review Summary section contains four tabs:

- Prioritized Review below
- Coverage Review on the next page
- Project Validation on page 58
- Manually-Selected on page 60

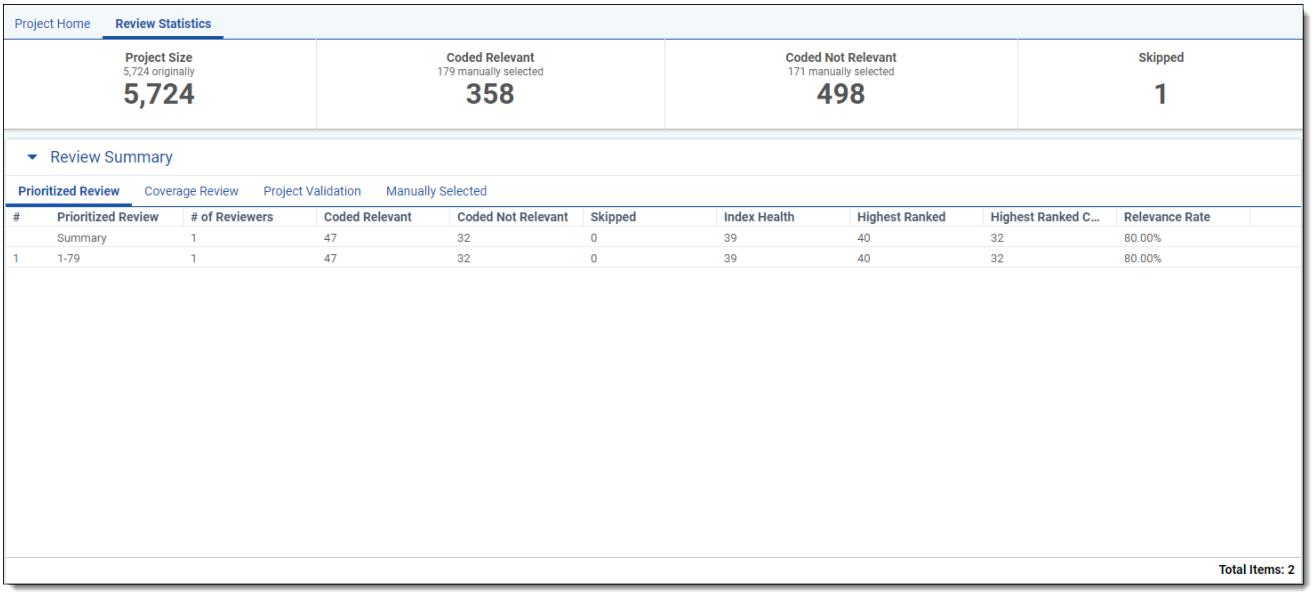
7.1.1 Prioritized Review

The Prioritized Review tab shows the effectiveness of the Prioritized Review queue's ability to locate relevant documents by reporting the review field breakdown and relevance rate each 200 documents. For every 200 documents that are coded (excluding additional reviewed by family), a new row appears in the Prioritized Review table. The first row in the table provides a summary for the entire project.

The Prioritized Review table contains the following columns:

- **Prioritized Review** - the set of documents the statistics apply to (excludes additional reviewed by family). The sum of the count of **Coded [Positive Choice]**, **Coded [Negative Choice]**, and **Skipped** documents should equal 200.
- **# of Reviewers** - the number of unique reviewers who reviewed documents in the Prioritized Review queue during this interval.
- **Coded [Positive Choice]** - the number of documents coded with the positive designation on the review field (excludes additional reviewed by family).
- **Coded [Negative Choice]** - the number of documents coded with the negative designation on the review field (excludes additional reviewed by family).
- **Skipped** - the number of documents that were saved or had **Save and Next** selected with no coding decision supplied on the review field (excludes additional reviewed by family).
- **Index Health** - the number of index health documents reviewed in the Prioritized Review queue. These documents are excluded from the relevance rate calculation.
- **Highest Ranked** - the number of highly ranked documents reviewed in the Prioritized Review queue.
- **Highest Ranked Coded [Positive Choice]** - the number of highly ranked documents that were coded with the positive designation in the Prioritized Review queue.
- **Relevance Rate** - the percentage of documents that were chosen for being highly ranked that were then confirmed as relevant by reviewers' coding decisions. You can calculate the relevance rate manually using the following formula: Highest Ranked Coded [Positive Choice] / Highest Ranked.
- **Additional Reviewed by Family** - the number of family documents reviewed in the Prioritized Review queue.

- **Additional Review by Family Coded [Positive Choice]** - the number of family documents coded with the positive designation in the Prioritized Review queue.

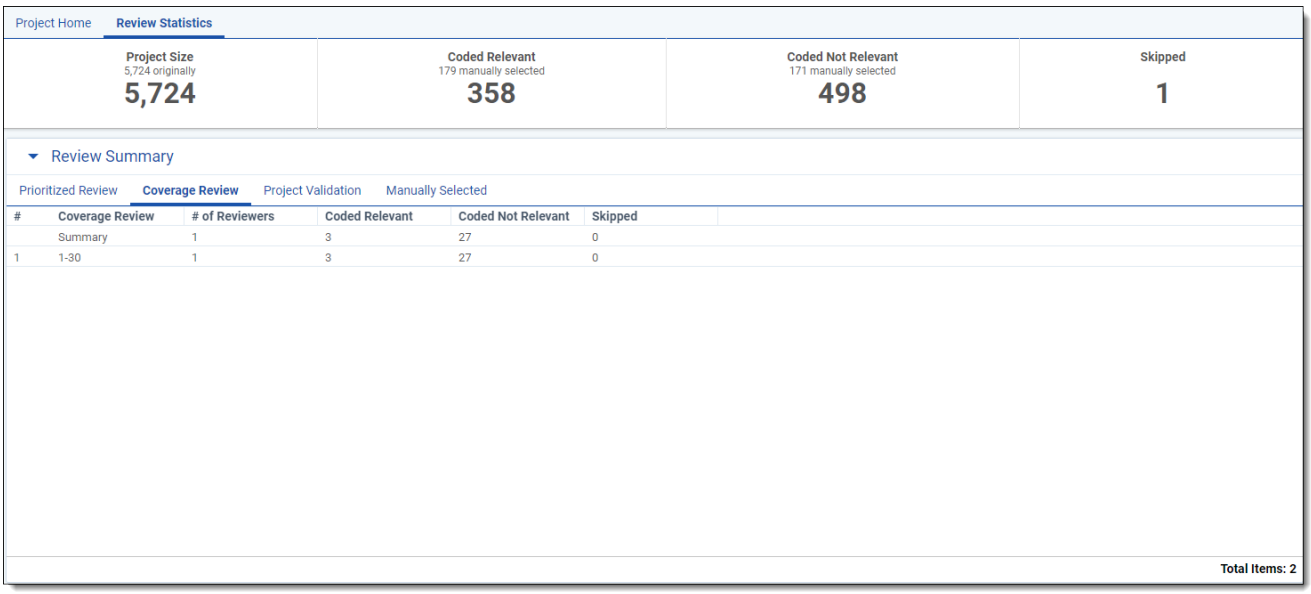


7.1.2 Coverage Review

The Coverage Review tab shows the progress of the Coverage Review by reporting the review field breakdown every 200 documents. For every 200 documents that are coded, a new row appears in the Coverage Review table. The first row in the table provides a summary for the entire project.

The Coverage Review table contains the following columns:

- **Coverage Review** - the set of documents the statistics apply to.
- **# of Reviewers** - the number of unique reviewers who reviewed documents in the Coverage Review queue during this interval..
- **Coded [Positive Choice]** - the number of documents coded with the positive designation on the review field.
- **Coded [Negative Choice]** - the number of documents coded with the negative designation on the review field.
- **Skipped** - the number of documents that were saved or had **Save and Next** selected with no coding decision supplied on the review field.



7.1.3 Project Validation

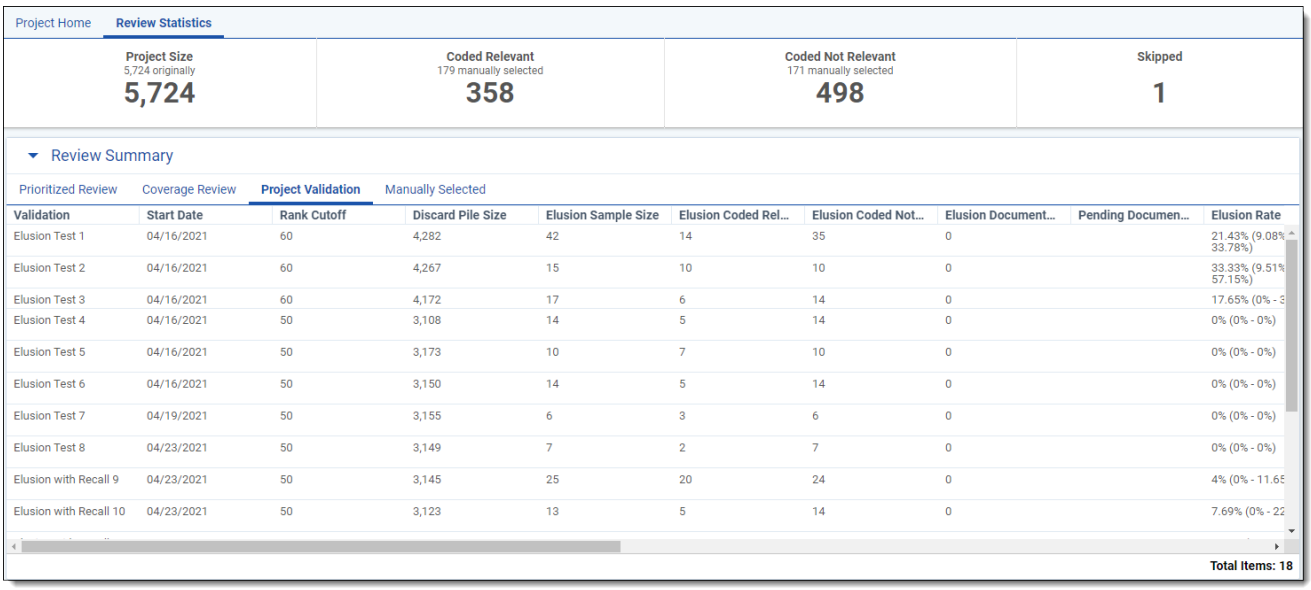
Use the Project Validation tab to monitor validation testing. A new row is created when Project Validation starts and is populated with the available information about the validation. As documents are coded, the counts and calculations update upon page refresh. If Project Validation is stopped before completion, statistics are calculated based on what has been coded so far.

The Project Validation tab contains the following columns:

- **Validation** - each review is called Elusion with Recall or Elusion Test plus a numeral. For example, the first Elusion with Recall is "Elusion with Recall 1," and the second is "Elusion with Recall 2."
- **Start Date** - the UTC date and time when Project Validation was started.
- **Rank Cutoff** - the numeric cutoff for positive prediction, fixed before project validation begins.
- **Discard Pile Size** - the number of documents below the rank cutoff that are not coded when Project Validation was started.
- **Elusion Sample Size** - the number of documents sampled for elusion rate. This number is computed when Project Validation is started.
- **Elusion Coded [Positive Choice]** - the number of sampled documents from below the cutoff which were coded positive during project validation.
- **Elusion Coded [Negative Choice]** - the number of sampled documents from below the cutoff which were coded negative during project validation.
- **Elusion Documents Skipped** - the number of sampled documents from below the cutoff that were saved or had **Save and Next** selected with no coding decision supplied on the review field.

Note: For best results, we strongly recommend coding every document in the Project Validation queue and avoiding skipping documents. For more information, see How Project Validation handles skipped documents on page 55.

- **Pending Document Count** - the number of documents whose coding has changed since the last model build (prior to Project Validation starting). This includes documents coded in Project Validation, and documents coded through other means. For instance, if a reviewer manually codes documents after Project Validation is underway, these will be counted as Pending.
- **Elusion Rate (Range)** - the error rate of the discard pile. This is calculated as the percentage of sampled, previously uncoded documents from below the cutoff which are coded positive during Project Validation. The range applies this sampled rate to the entire discard pile, using the confidence level provided by the user and the margin of error calculated from sample size.
- **Confidence Level for Elusion** - user input or calculated when setting up Project Validation.
- **Elusion Margin of Error** - margin of error calculated based on the elusion sample size, the discard pile size, and the elusion rate on the validation sample.
- **Estimated Eluded Documents (Range)** - the projected number of eluded documents. This estimates the number of relevant documents you would miss if you produced all documents marked relevant, as well as those with ranks at or above the cutoff.
- **Recall Rate (Range at CL80%)** - the percentage of truly positive documents which were found by the Active Learning process. A document has been "found" if it was previously coded positive, or if it is uncoded with a rank at or above the cutoff. Recall is calculated on the sample, then estimated for the total population with a confidence level (CL) of 80%.
- **Precision Rate (Range)** - the percentage of found documents which are truly positive. A document has been “found” if it was previously coded positive, or if it is uncoded with a rank at or above the cutoff. Documents which were predicted positive but coded negative during validation will count against precision. Precision is calculated on the sample, then estimated for the total population with a confidence level (CL) of 80%.
- **Precision Margin of Error (CL80%)** - the margin of error for precision as estimated from the sample size, the equivalent portion of the whole project, and the observed precision rate on the validation sample.
- **Richness Rate (Range)** - the percentage of documents which are relevant (positive choice). This is calculated by dividing the number of positive-coded documents in the sample by the total number of documents in the sample. The range predicts the richness for the whole project, subject to a 95% confidence level.
- **Richness Margin of Error (CL95%)** - the margin of error for richness as estimated from the sample size, the whole project size, and the observed richness rate on the sample.
- **Estimated Total Relevant Documents** - the estimated number of relevant documents in the whole project. This is calculated by projecting the richness rate across the whole project.
- **Total Documents in Project** - the number of documents in the project at the time of project validation.



Validation	Start Date	Rank Cutoff	Discard Pile Size	Elusion Sample Size	Elusion Coded Rel...	Elusion Coded Not...	Elusion Document...	Pending Documen...	Elusion Rate
Elusion Test 1	04/16/2021	60	4,282	42	14	35	0		21.43% (9.08% 33.78%)
Elusion Test 2	04/16/2021	60	4,267	15	10	10	0		33.33% (9.51% 57.15%)
Elusion Test 3	04/16/2021	60	4,172	17	6	14	0		17.65% (0% - 3
Elusion Test 4	04/16/2021	50	3,108	14	5	14	0		0% (0% - 0%)
Elusion Test 5	04/16/2021	50	3,173	10	7	10	0		0% (0% - 0%)
Elusion Test 6	04/16/2021	50	3,150	14	5	14	0		0% (0% - 0%)
Elusion Test 7	04/19/2021	50	3,155	6	3	6	0		0% (0% - 0%)
Elusion Test 8	04/23/2021	50	3,149	7	2	7	0		0% (0% - 0%)
Elusion with Recall 9	04/23/2021	50	3,145	25	20	24	0		4% (0% - 11.65
Elusion with Recall 10	04/23/2021	50	3,123	13	5	14	0		7.69% (0% - 22

7.1.4 Manually-Selected

The Manually-selected Documents chart displays the number of document coding decisions made outside of the Active Learning queue grouped by date.

The Manually-Selected tab contains the following columns:

- **Manually-selected Documents** - the number of documents coded outside of the Active Learning queue.
- **Coded [Positive Choice]** - the number of these documents coded with the positive designation on the review field.
- **Coded [Negative Choice]** - the number of these documents coded with the negative designation on the review field.
- **Date submitted (UTC)** - the date in UTC that the statistics were submitted.

Project Home | Review Statistics

Project Size	Coded Relevant	Coded Not Relevant	Skipped
5,724 originally	179 manually selected	171 manually selected	
5,724	358	498	1

Review Summary

Prioritized Review | Coverage Review | Project Validation | Manually Selected

#	Manually Selected	Coded Relevant	Coded Not Relevant	Date Submitted
1	245	139	106	04/16/2021
2	10	4	6	05/06/2021
3	95	36	59	05/07/2021

Total Items: 3

7.2 Model Updates

The Model Updates section contains a history of Active Learning model builds. A new row is added each time the model builds, and the statistics are based off of the <Positive Choice> Cutoff set in the Project Settings. If you update the <Positive Choice> Cutoff at any point, the statistics will update accordingly.

- **Build Date** - a timestamp indicating when the model build completed, displayed in local time.
- **Above or At Cutoff** - the number of documents above or at the <Positive Choice> Cutoff.
- **Below Cutoff** - the number of documents below the <Positive Choice> Cutoff.

The subsequent columns indicate the number of documents in each relevance rank range.

Model Updates

10 per page | 1 / 27

#	Build Date	Above or At Cutoff...	Below Cutoff (50)	0 - 10	11 - 20	21 - 30	31 - 40	41 - 50	51 - 60	61 - 70	71 - 80	81 - 90	91 - 100
1	06/04/2021 11:42 AM ...	2322	3402	30	216	1247	1480	463	701	991	591	5	0
2	05/07/2021 9:09 AM C...	2320	3404	47	198	1247	1484	458	695	994	597	4	0
3	05/07/2021 8:18 AM C...	2321	3403	47	198	1239	1484	462	704	990	595	5	0
4	05/07/2021 7:58 AM C...	2302	3422	22	217	1195	1516	505	763	975	527	4	0
5	05/07/2021 7:03 AM C...	2301	3423	22	218	1171	1535	513	778	974	510	3	0
6	05/06/2021 11:29 AM ...	2296	3428	7	222	1132	1565	545	818	953	479	3	0
7	05/05/2021 1:29 PM C...	2305	3419	3	225	1102	1579	553	823	960	477	2	0
8	05/04/2021 3:05 PM C...	2308	3416	3	218	1087	1579	571	817	975	472	2	0
9	05/04/2021 2:43 PM C...	2315	3409	5	216	1025	1612	595	813	991	465	2	0
10	05/04/2021 2:18 PM C...	2310	3414	2	202	924	1688	640	904	967	397	0	0

8 Active Learning performance baselines

This page is meant to be used as a reference to track the overall performance of Active Learning in RelativityOne. It should not be used as a benchmark of what you expect to see in a production client environment or Relativity Server environment due to differences in data, infrastructure, and configuration. The results may not scale linearly. Exceeding these limits may result in failure or degraded experience using Active Learning.

8.1 Active Learning project size recommendations

These recommendations are the result of extensive testing in RelativityOne. For the best user experience, we advise adding a maximum of 9 million total documents and a maximum of 1 million coded documents to an Active Learning project .

Note: We recommend no more than 150 concurrent reviewers per project. Concurrent reviewers are defined as reviewers making coding decisions in an Active Learning queue at the same time. There is no limit to how many reviewers you can add to a queue as long as the number of concurrent reviewers remains at 150 or fewer.

Max documents in classification index	Max coded documents	Max concurrent reviewers
9 million	1 million	150

8.2 Data set details

These tests were run on a subset of the following data set in a RelativityOne environment. Results may vary.

Data set name	File count	Average extracted text size (KB)	Total Extracted Text size (GB)
Real World	9,154,516	30.18	276.25

8.3 Classification index population + build results

A classification index is required for an Active Learning project. This performance run includes population of all documents and building. Start time was measured as the time the first document was sent to the Analytics server, and end time was measured as when the last document became available in Active Learning.

Index size (Documents)	Pre-coded documents	Population rate (GB/hr)	Population time (h:mm:ss)	Index build (h:mm:ss)	Total operation time (h:mm:ss)	Documents/hr
1,000,000	10	13	3:03:29	0:13:30	3:16:59	304,878

8.4 Active Learning index build results

Once the Active Learning model completes its initial build, the model builds at maximum every 20 minutes to include new coded documents. The documents were randomly coded 50% responsive and 50% non-responsive using Relativity's sampling feature.

The test scenarios in the following table use an index that contained 1,000,000 documents.

Note: The results listed below were measured after the initial model build completed and do not include the population stage as the documents have already been added to the index.

	Coded documents	Index build time (h:mm:ss)
Build 1	10	0:13:30
Build 2	100,000	0:27:17
Build 3	200,000	0:31:02
Build 4	300,000	0:38:29
Build 5	400,000	0:50:52
Build 6	500,000	0:56:21
Build 7	600,000	1:06:27
Build 8	700,000	1:09:29
Build 9	800,000	1:09:33
Build 10	900,000	2:45:09
Build 11	1,000,000	3:24:16

The test scenarios in the following table use an index that contained 9,154,516 documents.

	Coded documents	Index build time (h:mm:ss)
Build 1	10	1:38:30
Build 2	100,000	2:59:32
Build 3	700,000	5:14:47
Build 4	800,000	7:39:47
Build 5	900,000	6:50:19
Build 6	1,000,000	6:25:35

8.5 Update ranks results

In an Active Learning project, you can manually update the document ranks and ensure the rank categorization field is up to date. Once you click Update Ranks, you can monitor the update progress via a fly-out modal. You can update ranks again only after the current modification is complete. Update ranks performs faster on subsequent updates than on the initial update.

Project size (documents)	Initial update ranks Operation time (h:mm:ss)	Subsequent update ranks Operation time (h:mm:ss)
200,000	0:02:48	0:01:59
400,000	0:06:12	0:04:42
600,000	0:09:42	0:07:02
800,000	0:13:48	0:11:39
1,000,000	0:22:24	0:15:14
9,154,516	5:47:36	3:39:54

Proprietary Rights

This documentation ("**Documentation**") and the software to which it relates ("**Software**") belongs to Relativity ODA LLC and/or Relativity's third party software vendors. Relativity grants written license agreements which contain restrictions. All parties accessing the Documentation or Software must: respect proprietary rights of Relativity and third parties; comply with your organization's license agreement, including but not limited to license restrictions on use, copying, modifications, reverse engineering, and derivative products; and refrain from any misuse or misappropriation of this Documentation or Software in whole or in part. The Software and Documentation is protected by the **Copyright Act of 1976**, as amended, and the Software code is protected by the **Illinois Trade Secrets Act**. Violations can involve substantial civil liabilities, exemplary damages, and criminal penalties, including fines and possible imprisonment.

©2024. Relativity ODA LLC. All rights reserved. Relativity® is a registered trademark of Relativity ODA LLC.