# Content Analyzer Plug-in for GAX

## Testing Models

12/14/2025

# Contents

# Testing Models

Once you've created a model, you can test it. Testing takes a model and has it analyze a training object. A moment's thought will tell you that the training object must be

- One with the same root category as the model
- Not the one that was used to create the model

Schedule the test on the Testing Schedule tab: simply select the model, the TDO to test it on, and the start time for the test.

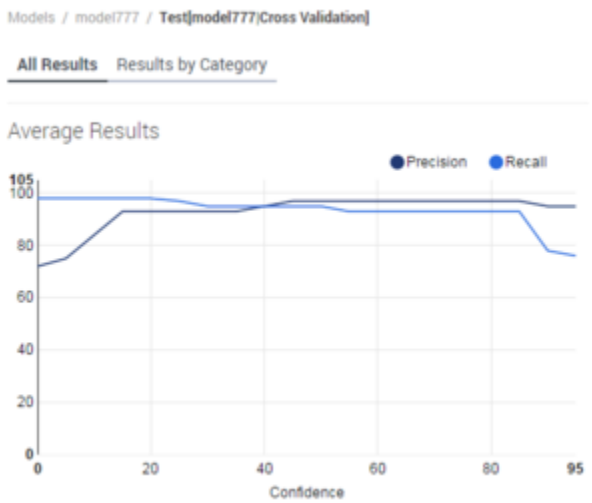
To see the test results for a model, select the model, on either the Testing Schedule or Models tab, and click the eye icon. The eye icon will be active only if you've select a model that has been tested.

Understanding the test results is where it gets interesting.

## All Results tab

This tab shows the Average Results and Correct in Top N graphs, and the Category Confusion table.

### Average Results




Precision, Recall, and Confidence

This graph shows the Precision (black) and Recall (blue) ratings (vertical axis) at a given Confidence level (horizontal axis). But what do those terms mean? Read on:

### Confidence

This is a numerical score, from 1 to 100, that indicates the percent likelihood, according to the selected model, that a text object belongs in a certain category.

(In contrast, *accuracy* is an assessment, produced by testing, of the correctness of a model's assignment of text objects to categories. In other words, confidence expresses a model's guess about a categorization; accuracy rates the correctness of that guess)

### Precision and Recall

To understand Precision and Recall, consider several possible ways of looking at the performance of a model. If your model attempts to assign a certain number of items to a category X, you can make the following counts:

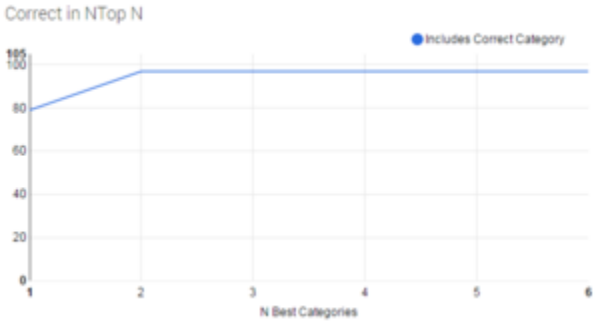
- *a* = the number of items the model correctly assigns to X
- *b* = the number of items the model incorrectly assigns to X
- *c* = the number of items the model incorrectly rejects from X (that is, items that the model should assign to X but does not)

From these quantities, you can calculate the following performance measures:

- Precision = $a/(a + b)$
- Recall = $a/(a + c)$

Generally, for increasing precision you pay the price of decreasing recall. That is, the model assigns an item to a category only when it is very sure that the item belongs. But by insisting on being very sure, it runs the risk of rejecting items that really do belong in the category. In the figure **Precision, Recall, and Confidence**, you can see this effect above the 85 percent Confidence level.

## Correct in Top N




Correct in Top N

When a model classifies a text object, it returns a list of categories and the probability (the Confidence rating) that the object belongs to them. Ranking the returned categories with the highest probability first, how likely is it that the correct category appears within the top two, the top three, and so on?

- The horizontal axis, **N Best Categories**, shows the best (top-ranked) category, best two, best three, and so on.
- The vertical axis, **Includes Correct Category**, is the percent likelihood that this many top-ranked categories include the correct one.

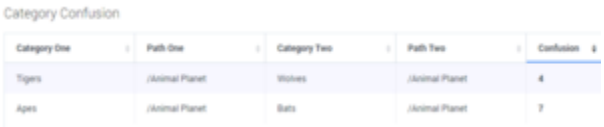
Imagine that the highest-probability returned category is the model's first guess. If we go down to the fourth-ranked category, that's like giving the model four guesses. The more guesses, the easier it is to get the right answer. The fewer guesses, the better the model is at classifying.

The model in the figure is quite good: its first guess (top-ranked returned category) is right just under 80 percent of the time, and we only have to give it one more guess to achieve close to 100 percent correct.

One way to use this rating is to advise agents how many categories to look at when choosing a standard response. If there is a 95 percent probability that the right category is in the top three, you can advise agents to consider only the top three categories.

## Category Confusion

The Category Confusion table lists up to 10 pairs of categories that the model is likely to confuse.




Category Confusion

The Confusion column shows the probability, as a percentage, that the model will mistakenly classify a Category 1 item as Category 2. For example, the figure **Category Confusion** shows that this model classifies tigers as wolves 4 percent of the time.

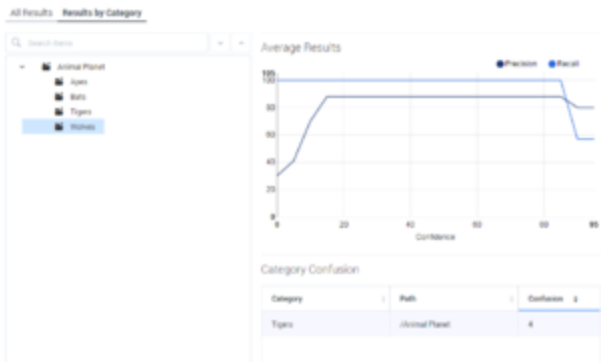
A rating of 50 would mean total confusion: the model cannot distinguish wolves from tigers. A rating of 100 would mean that the model always calls wolves tigers and always calls tigers wolves—a complete reversal.

If a pair of categories has a rating of over 20 and both categories have more than three or four members, you should consider modifying them. You can modify them in either of two directions:

- Merge them; that is, decide that they are so similar that they amount to a single category.
- Further differentiate them by adding more highly contrasting training interactions to them in the Training Data Object.

# Results by Category tab

This tab displays the same ratings as the **All Results** tab, but for a single category.




Results by Category

**Category Confusion** shows the category/ies that are likely to be confused with the category that's selected on the left.