Evaluation Metrics

Anote Presented by Harsh Thakkar

Retrieval Accuracy

Getting correct evidence text for answer

Answering Accuracy

Getting correct answer

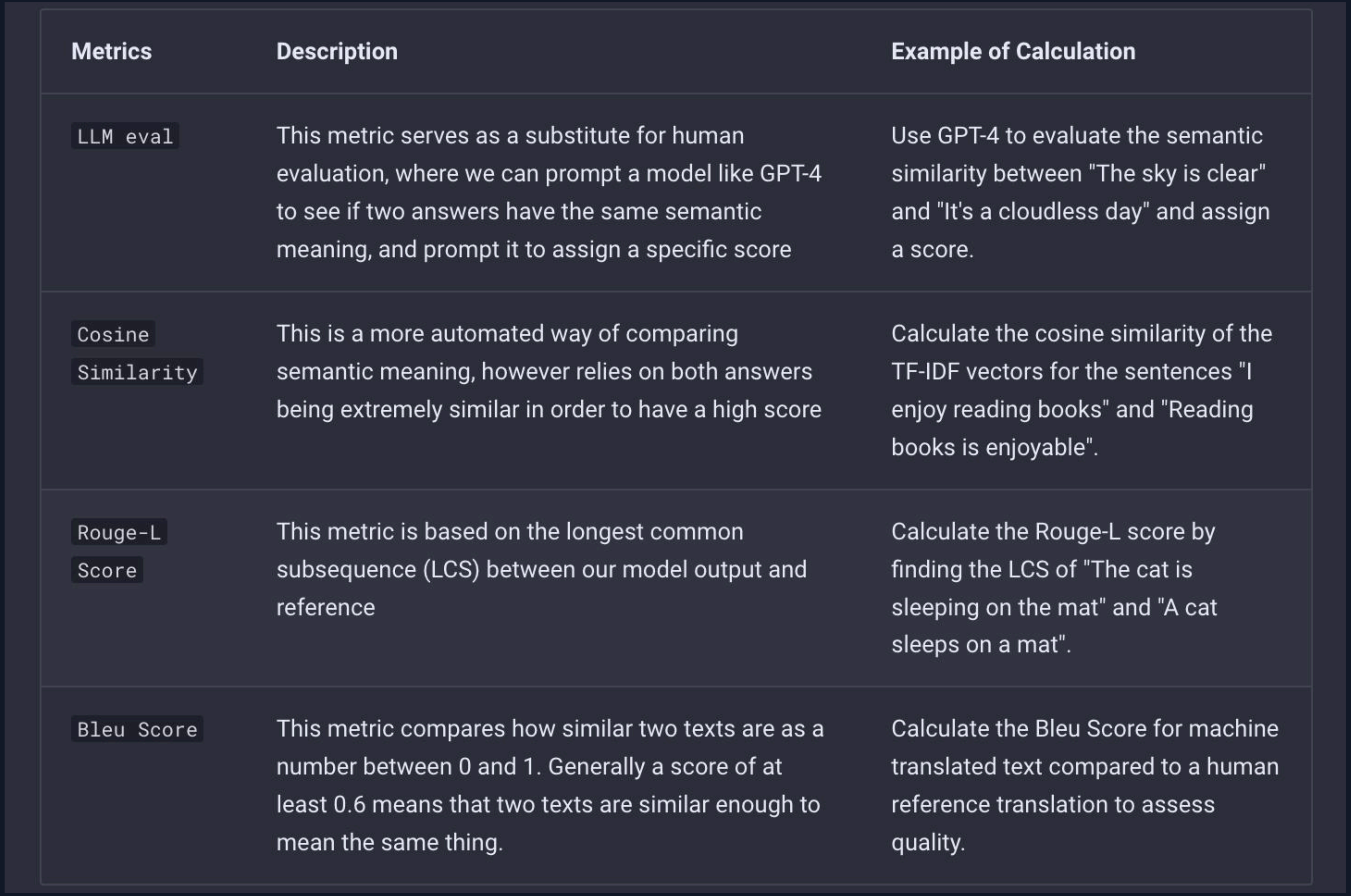
Structured Metrics

Evaluating vs. Ground Truth Label

Unstructured Metrics

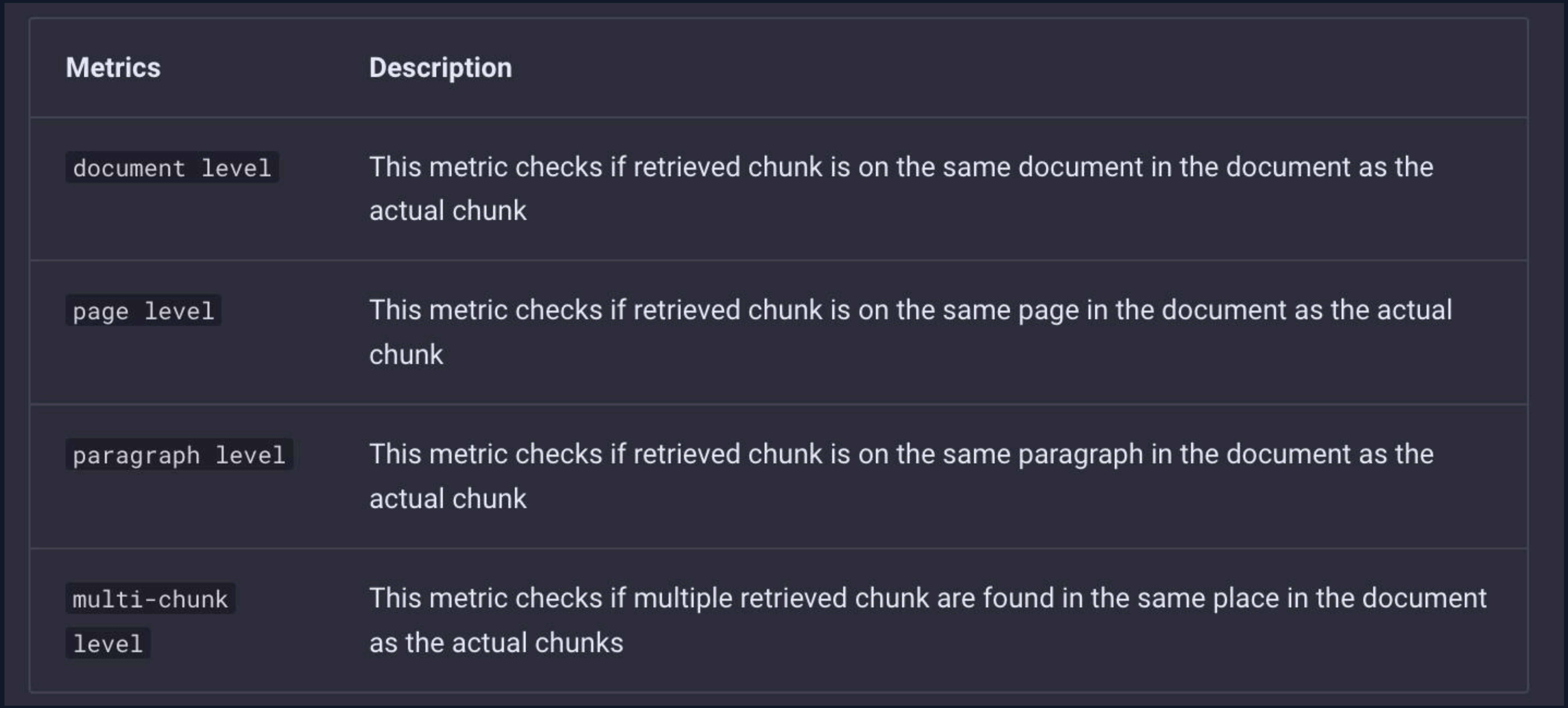
Evaluating without Ground Truth Label

Structured Answer Accuracy Metrics



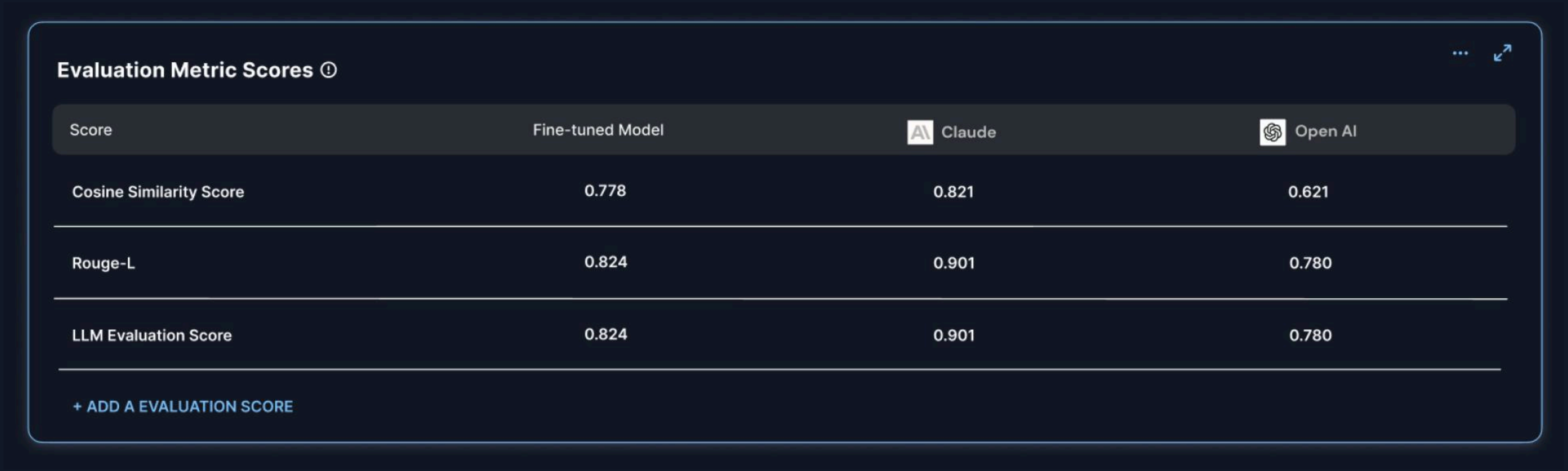
Metrics	Description	Example of Calculation
`LLM eval`	This metric serves as a substitute for human evaluation, where we can prompt a model like GPT-4 to see if two answers have the same semantic meaning, and prompt it to assign a specific score	Use GPT-4 to evaluate the semantic similarity between "The sky is clear" and "It's a cloudless day" and assign a score.
`Cosine Similarity`	This is a more automated way of comparing semantic meaning, however relies on both answers being extremely similar in order to have a high score	Calculate the cosine similarity of the TF-IDF vectors for the sentences "I enjoy reading books" and "Reading books is enjoyable".
`Rouge-L Score`	This metric is based on the longest common subsequence (LCS) between our model output and reference	Calculate the Rouge-L score by finding the LCS of "The cat is sleeping on the mat" and "A cat sleeps on a mat".
`Bleu Score`	This metric compares how similar two texts are as a number between 0 and 1. Generally a score of at least 0.6 means that two texts are similar enough to mean the same thing.	Calculate the Bleu Score for machine translated text compared to a human reference translation to assess quality.

Structured Retrieval Metrics



Metrics	Description
`document level`	This metric checks if retrieved chunk is on the same document in the document as the actual chunk
`page level`	This metric checks if retrieved chunk is on the same page in the document as the actual chunk
`paragraph level`	This metric checks if retrieved chunk is on the same paragraph in the document as the actual chunk
`multi-chunk level`	This metric checks if multiple retrieved chunk are found in the same place in the document as the actual chunks

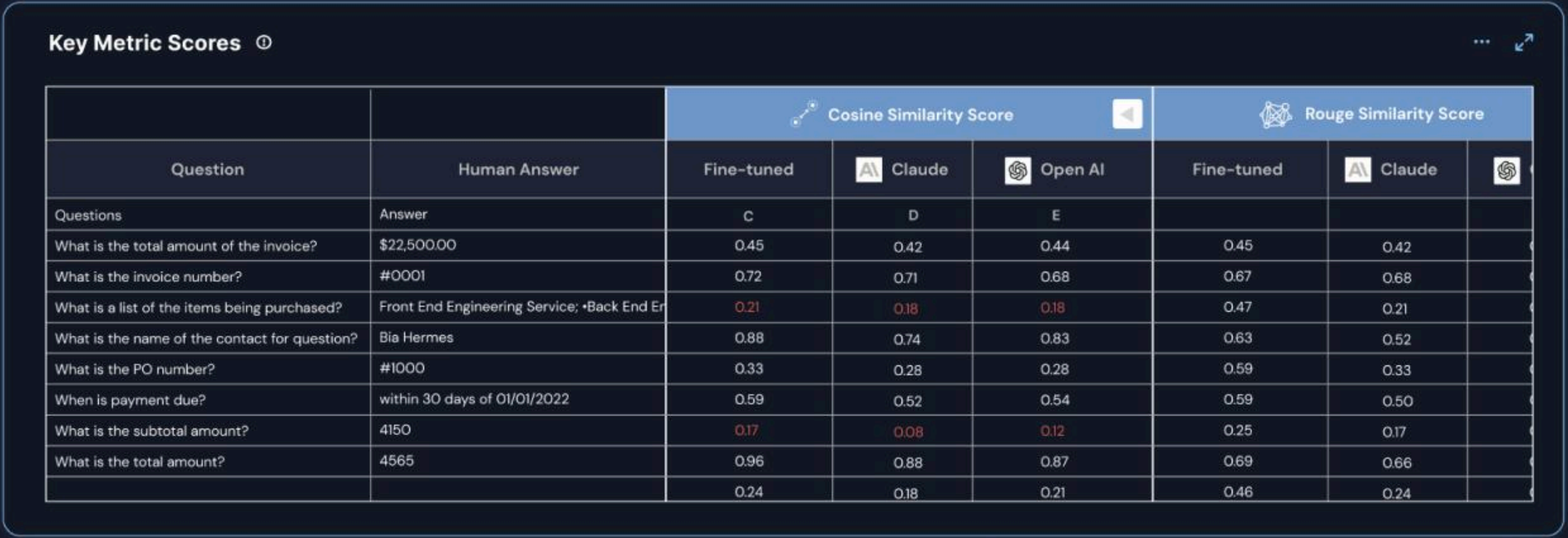
Aggregate Metrics



Evaluation Metric Scores

Score	Fine-tuned Model	Claude	Open AI
Cosine Similarity Score	0.778	0.821	0.621
Rouge-L	0.824	0.901	0.780
LLM Evaluation Score	0.824	0.901	0.780

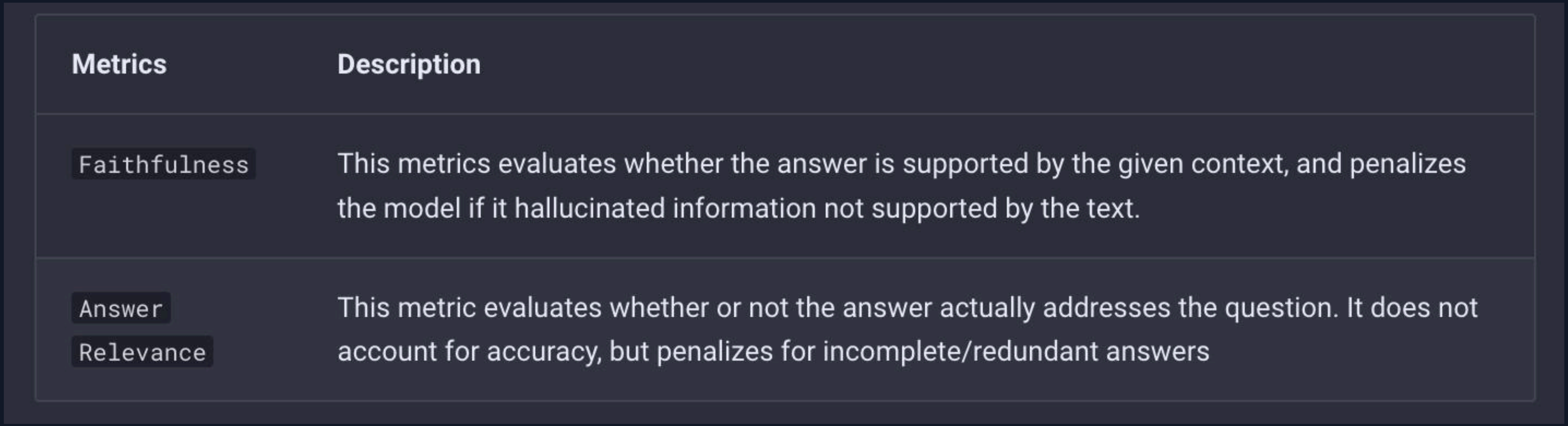
Row Specific Metrics



Key Metric Scores

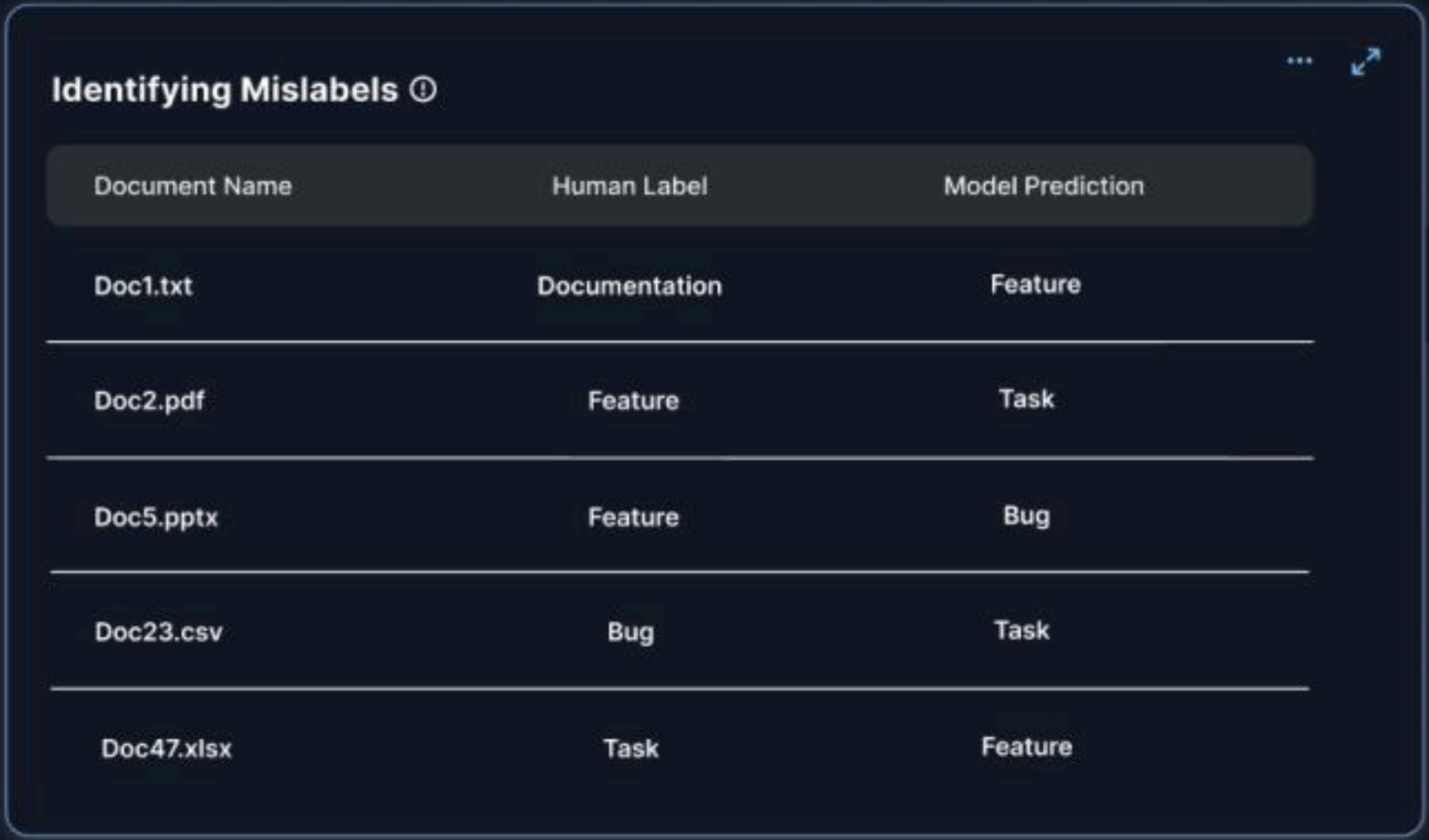
		Cosine Similarity Score			Rouge Similarity Score	
Question	Human Answer	Fine-tuned	Claude	Open AI	Fine-tuned	Claude
Questions	Answer	C	D	E		
What is the total amount of the invoice?	$22,500.00	0.45	0.42	0.44	0.45	0.42
What is the invoice number?	#0001	0.72	0.71	0.68	0.67	0.68
What is a list of the items being purchased?	Front End Engineering Service; •Back End Er	0.21	0.18	0.18	0.47	0.21
What is the name of the contact for question?	Bia Hermes	0.88	0.74	0.83	0.63	0.52
What is the PO number?	#1000	0.33	0.28	0.28	0.59	0.33
When is payment due?	within 30 days of 01/01/2022	0.59	0.52	0.54	0.59	0.50
What is the subtotal amount?	4150	0.17	0.08	0.12	0.25	0.17
What is the total amount?	4565	0.96	0.88	0.87	0.69	0.66
		0.24	0.18	0.21	0.46	0.24

Unstructured Answer Accuracy Metrics



Metrics	Description
`Faithfulness`	This metrics evaluates whether the answer is supported by the given context, and penalizes the model if it hallucinated information not supported by the text.
`Answer Relevance`	This metric evaluates whether or not the answer actually addresses the question. It does not account for accuracy, but penalizes for incomplete/redundant answers

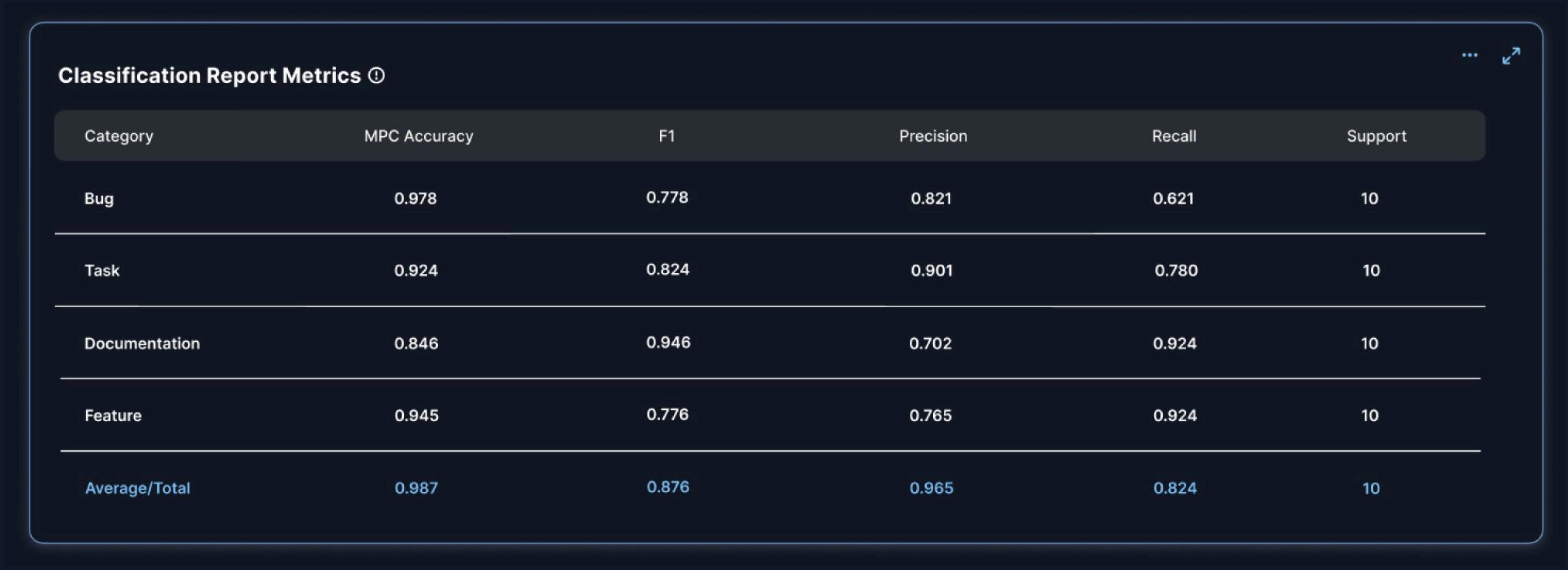
Identifying Mislabels



Identifying Mislabels ⓘ

Document Name	Human Label	Model Prediction
Doc1.txt	Documentation	Feature
Doc2.pdf	Feature	Task
Doc5.pptx	Feature	Bug
Doc23.csv	Bug	Task
Doc47.xlsx	Task	Feature

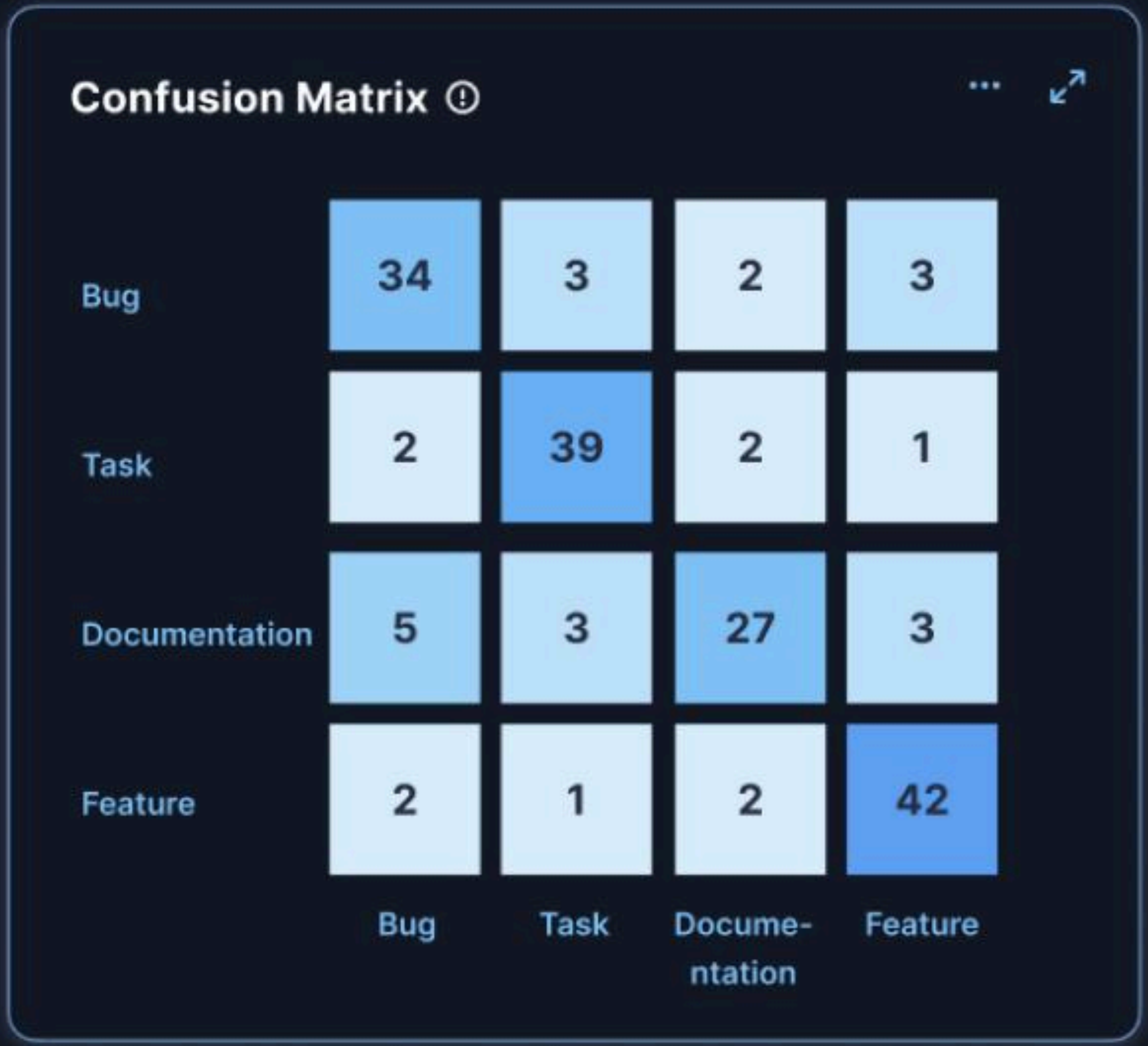
Classification Report



Classification Report Metrics ⓘ

Category	MPC Accuracy	F1	Precision	Recall	Support
Bug	0.978	0.778	0.821	0.621	10
Task	0.924	0.824	0.901	0.780	10
Documentation	0.846	0.946	0.702	0.924	10
Feature	0.945	0.776	0.765	0.924	10
Average/Total	0.987	0.876	0.965	0.824	10

Confusion Matrix



Confusion Matrix

	Bug	Task	Documentation	Feature
Bug	34	3	2	3
Task	2	39	2	1
Documentation	5	3	27	3
Feature	2	1	2	42

Product Demo