

# New Evaluation

Evaluate batch experiment results against ground truth labels.

## Navigation

Dashboard

Tables

Ontology

Labels

Batches

## Annotations

New Run

Run History

## Evaluation

+ New Evaluation

Compare

## Evaluation Setup

Select a batch result file and ground truth labels to evaluate.

### Batch Result

single | cot | BEO (431 cols) litellm azure-gpt-4.1

single cot BEO litellm azure-gpt-4.1

### Ground Truth Labels

Ground Truth Small Table Based On Name (431 columns)

▷ Run Evaluation

## Results

Evaluation completed: batch\_20251205\_224711

Evaluated

431

of 431 columns

Skipped

0

no ground truth

### Path-Level Metrics

Precision  
37.43%

Recall  
34.57%

F1  
34.84%

### Node-Level Metrics

Precision  
46.33%

Recall  
40.26%

F1  
41.45%

Config: single | cot | max\_depth=2

Labels: ground\_truth\_small\_table\_based\_on\_name.csv