Navigation

Dashboard

Tables

Ontology

Labels

Batches

Annotations

New Run

Run History

Evaluation

New Evaluation

Compare

Settings

Compare Experiments

Compare evaluation results across multiple experiments side-by-side.

Comparison Setup

Select batch files and ground truth labels to compare.

Ground Truth Labels

Ground Truth Small Table Based On Name

Batch Files to Compare

single | cot | BEO 431c litellm azure-gpt-4.1

edm | cot | BEO 431c ollama gpt-oss:20b

Add Another

Compare Experiments

Comparison Results

2 of 2 experiments compared successfully

Best Scores: Path F1: **34.84%** Node F1: **41.45%**

Experiment	Provider	Model	Mode	Prompt	Path F1	Node F1	Columns	Status
batch_20251205_224711	litellm	azure-gpt-4.1	single	cot	34.84%	41.45%	431/431	Success
batch_20251205_224635	ollama	gpt-oss:20b	edm	cot	23.51%	29.63%	431/431	Success

batch_20251205_224711

litellm azure-gpt-4.1

single | cot

Path-Level
P: 37.43% R: 34.57% **F1: 34.84%**

Node-Level
P: 46.33% R: 40.26% **F1: 41.45%**

batch_20251205_224635

ollama gpt-oss:20b

edm | cot

Path-Level
P: 25.14% R: 24.25% **F1: 23.51%**

Node-Level
P: 32.44% R: 29.81% **F1: 29.63%**