[Evaluation] - Metric Config and Management

Research	Confident AI Blog - Resources to help teams stay confident in AI Blog - Langfuse Evaluating AI Agents - DeepLearning.AI
Design	Papago AI-Evaluation
Relevant PRDs	[Evaluation] - Create and manage Test Case for Agent Evaluation

Business Context

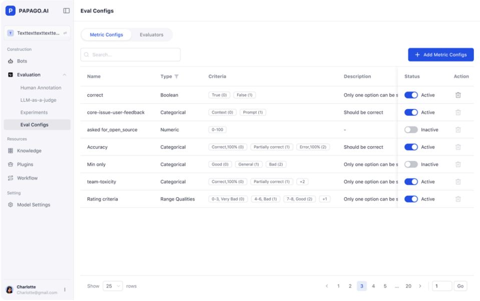
As an Agent Building Platform, our mission is to help users (agent builders) maximize their ROI from the effort invested in building and refining agents. Similar to traditional software development, having clear indicators that signal improvements after each update is crucial for product growth and success. In this phase of agent optimization, users seek solutions to construct a systematic evaluation workflow that provides consistent and reliable insights through metrics that matter most to them.

Requirement

1. Metrics FE-2346: [Eval Config] list of Metrics LAUNCHED

<table>
<tr><th>Requirement</th><th>Wireframe / Design</th></tr>
<tr><td>Add Eval Config:

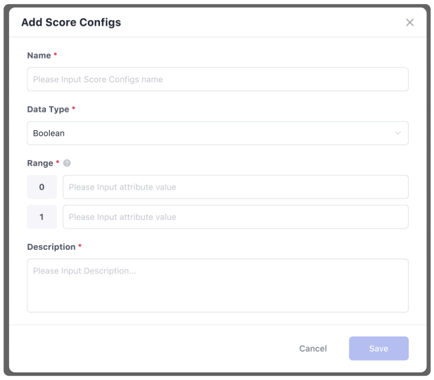
1. Metric Config
2. Evaluator

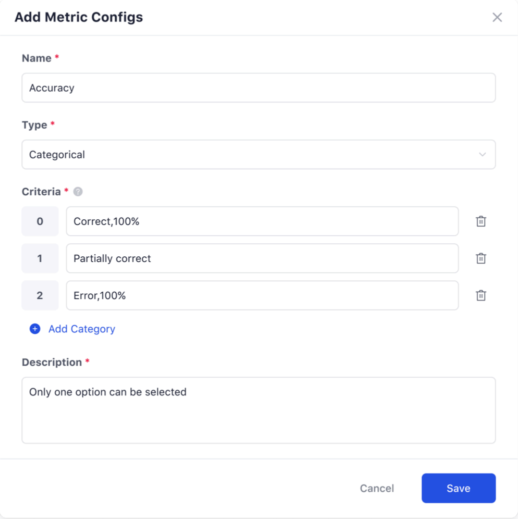
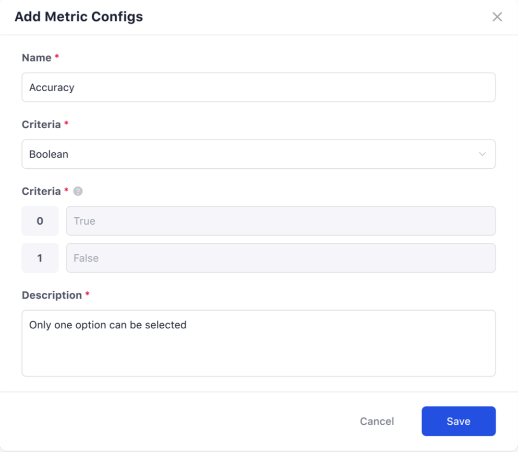
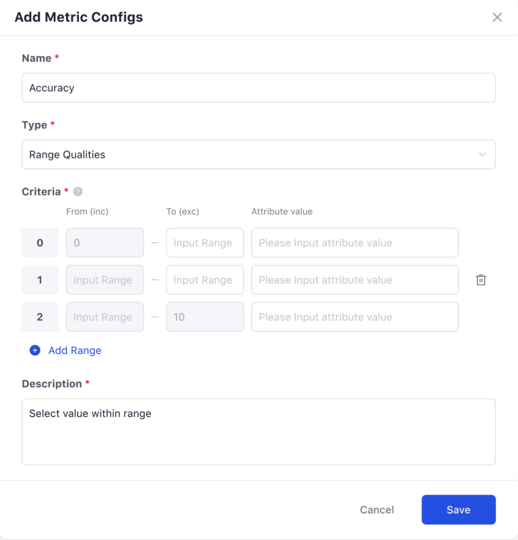
Metric tab show all metrics</td><td>

Name	Type	Criteria	Description	Status	Action
correct	Boolean	True (0) False (1)	Only one option can be s	Active	
core-issue-user-feedback	Categorical	Context (0) Prompt (1)	Should be correct	Active	
asked for_open_source	Numeric	0-100	-	Inactive	
Accuracy	Categorical	Correct,100% (0) Partially correct (1) Error,100% (2)	Should be correct	Active	
Min only	Categorical	Good (0) General (1) Bad (2)	Only one option can be s	Inactive	
team-toxicity	Categorical	Correct,100% (0) Partially correct (1) +2	Only one option can be s	Active	
Rating criteria	Range Qualities	0-3, Very Bad (0) 4-6, Bad (1) 7-8, Good (2) +1	Only one option can be s	Active	
</td></tr>
<tr><td>Metrics are searchable, filterable by data type, evaluators, creators</td><td></td></tr>
</table>

Metrics is also support pagination and adjust display number. Inherit behavior from existing listing	
Columns are immutable. Shown columns are: 1. Name: text 2. Type: selected type 3. Criteria: text, each criteria separate by comma `,` 4. Description: long text 5. Creator: Workflow member 6. Reference in (Run): Number of reference to Evaluator	
Switchable column: Status: Active, Draft	
Action include edit, remove	

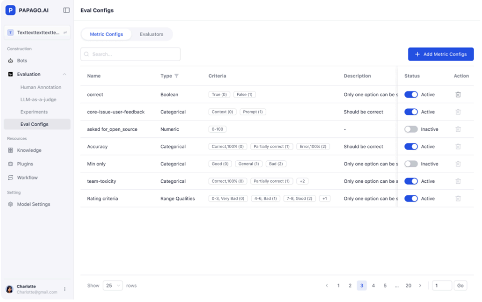
2. Create a new metric

Requirement	Wireframe / Design
User can create a new metric with: 1. Name 2. Type 3. Criteria (Range) 4. Description (Explain why the metric matter and how to measure the metric)	 Add Score Configs Name * Data Type * Boolean Range * 0 1 Description * Please Input Score Configs name Please Input attribute value Please Input attribute value Please Input Description... Cancel Save

1. Boolean type will only have two value **True, False**. Non editable. Allow editing this would create confusion in the future. 2. **Numeric** will have default 0-100 given this range is very common. 3. **Categorical** unchanged 4. Add a new type Range Qualities to allow users to config cases that have quantitative score to be map with its qualitative label: a. **0-100%:** 0% error, 25% bad, 50% normal, ... 100% <quality label> b. **1-7:** 1 label A, 2 label B, ...	 Add Metric Configs Name * Accuracy Type * Categorical Criteria * 0 Correct,100% 1 Partially correct 2 Error,100% Description * Only one option can be selected  Add Metric Configs Name * Accuracy Criteria * Boolean Criteria * 0 True 1 False Description * Only one option can be selected  Add Metric Configs Name * Accuracy Type * Range Qualities Criteria * From (inc) — To (exc) — Attribute value 0 0 — Input Range — Please Input attribute value 1 Input Range — Input Range — Please Input attribute value 2 Input Range — 10 — Please Input attribute value Description * Select value within range

3. Manage metrics

Requirement	Wireframe / Design

Metric can be switch from active to draft	
Metric can be deleted. There will be a confirmation pop-up when deleting a metric **Deleting or changing a metric will not affect the result of experiment runs or production audit run**	

Eval Configs

Name	Type	Criteria	Description	Status	Action
correct	Boolean	True (0) False (1)	Only one option can be s	Active	
core-issue-user-feedback	Categorical	Context (0) Prompt (1)	Should be correct	Active	
asked for_open_source	Numeric	0-100	-	Inactive	
Accuracy	Categorical	Correct,100% (0) Partially correct (1) Error,100% (2)	Should be correct	Active	
Min only	Categorical	Good (0) General (1) Bad (2)	Only one option can be s	Inactive	
team-toxicity	Categorical	Correct,100% (0) Partially correct (1) +2	Only one option can be s	Active	
Rating criteria	Range Qualities	0-3, Very Bad (0) 4-6, Bad (1) 7-8, Good (2) +1	Only one option can be s	Active	

Show 25 rows ‹ 1 2 3 4 5 ... 20 › 1 Go

4. Metric Usage

Metrics could be use in Evaluator configuration and will be in the scope of Evaluator PRD