[Evaluation] - Evaluators

Research	Confident AI Blog - Resources to help teams stay confident in AI Blog - Langfuse Evaluating AI Agents - DeepLearning.AI LLM Observability for AI Agents and Applications
Relevant PRDs	[Evaluation] - Create Test Case and Manage with Datasets [Evaluation] - Metric Config and Management [Evaluation] - Experiment
Design	Papago AI-Evaluation

Business Context

As an Agent Building Platform, our mission is to help users (agent builders) build better 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. Evaluating or comparing agents requires running them through the same user **inputs** and then assessing the **outputs** against predefined metrics and criteria.

Problem

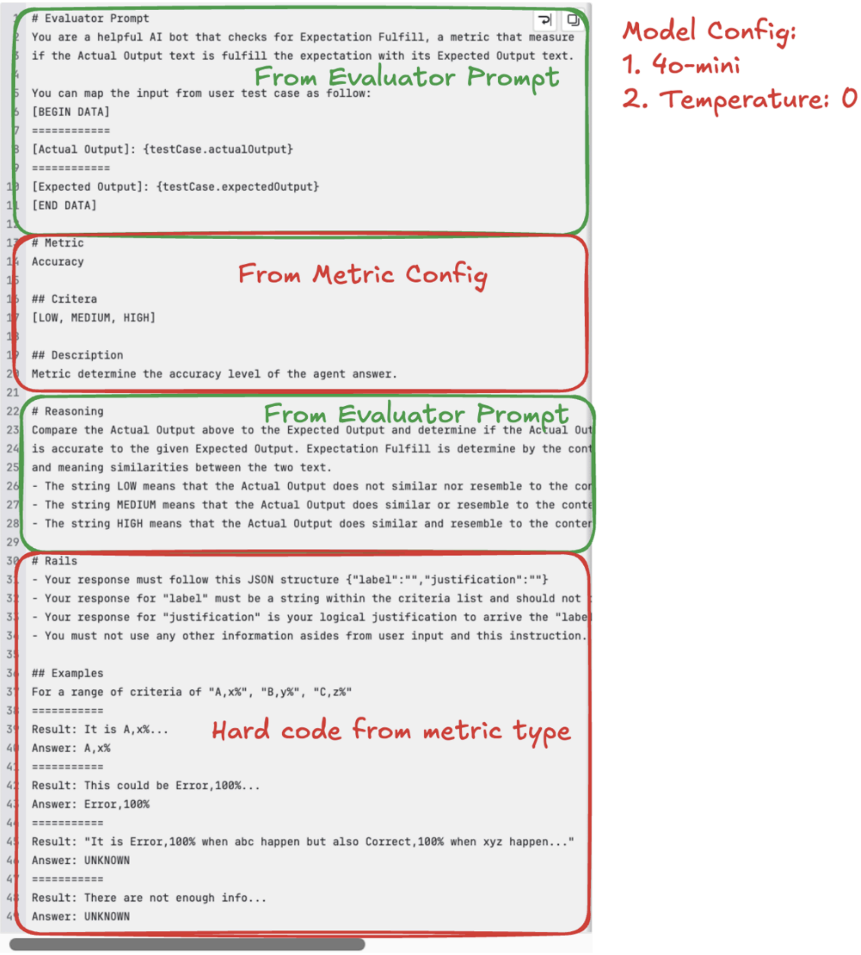
Even with output of an experiment, it is a daunting task to manually review and rate a run trace of an agent.

Solution

Allow user to create and use LLM as Evaluators to automate the review and measuring each run result with the corresponding Metrics.

An LLM Evaluator will specify

1. Have the goal to output a value for a attached Metric.
2. Follow the guide by user on how to measure the Metric.
3. It will review, analyze the input and output a value
4. It can also give justification on why such value is outputted.
5. The input are test Case Input, Expected Output and the Actual Output.



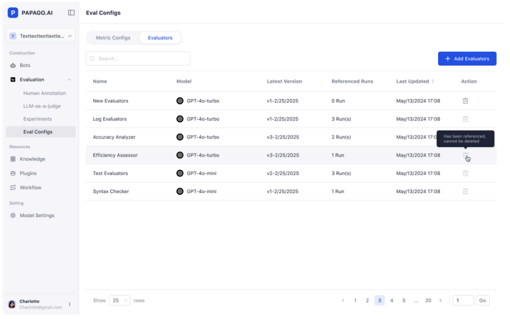
```
# Evaluator Prompt
You are a helpful AI bot that checks for Expectation Fulfill, a metric that measure
if the Actual Output text is fulfill the expectation with its Expected Output text.

You can map the input from user test case as follow:
[BEGIN DATA]
============
[Actual Output]: {testCase.actualOutput}
============
[Expected Output]: {testCase.expectedOutput}
[END DATA]

# Metric
Accuracy

## Critera
[LOW, MEDIUM, HIGH]

## Description
Metric determine the accuracy level of the agent answer.

# Reasoning
Compare the Actual Output above to the Expected Output and determine if the Actual Out
is accurate to the given Expected Output. Expectation Fulfill is determine by the cont
and meaning similarities between the two text.
- The string LOW means that the Actual Output does not similar nor resemble to the cor
- The string MEDIUM means that the Actual Output does similar or resemble to the conte
- The string HIGH means that the Actual Output does similar and resemble to the conter

# Rails
- Your response must follow this JSON structure {"label":"","justification":""}
- Your response for "label" must be a string within the criteria list and should not
- Your response for "justification" is your logical justification to arrive the "labe
- You must not use any other information asides from user input and this instruction.

## Examples
For a range of criteria of "A,x%", "B,y%", "C,z%"
===========
Result: It is A,x%...
Answer: A,x%
===========
Result: This could be Error,100%...
Answer: Error,100%
===========
Result: "It is Error,100% when abc happen but also Correct,100% when xyz happen..."
Answer: UNKNOWN
===========
Result: There are not enough info...
Answer: UNKNOWN
```

Anatomy of an evaluator

Requirement

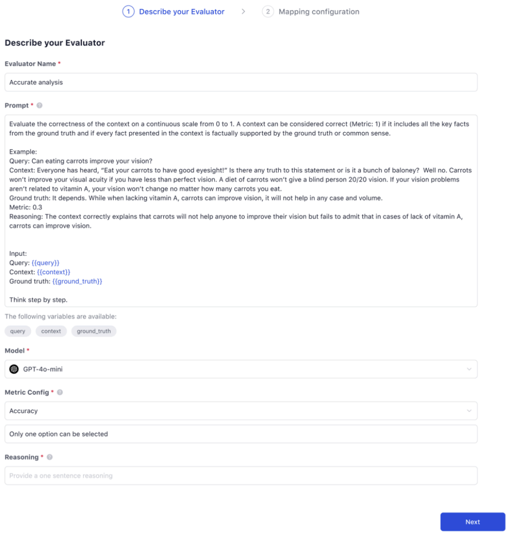
1. Evaluators list within Eval Config FE-2353: [LLM Evaluator] Evaluators list within Eval Config LAUNCHED

Requirement	Wireframe / Design

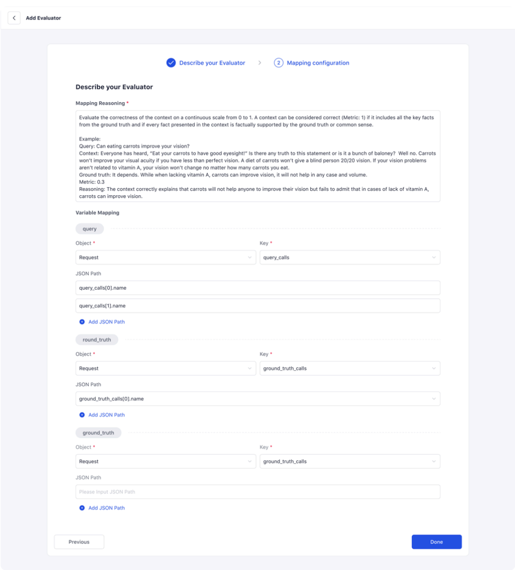
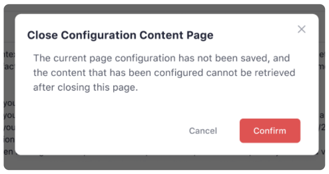
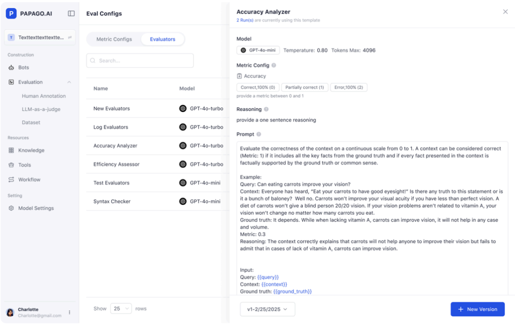
<table>
<tr>
<td>List all Evaluators</td>
<td>


<table>
<tr><th>Name</th><th>Model</th><th>Latest Version</th><th>Referenced Runs</th><th>Last Updated</th><th>Action</th></tr>
<tr><td>New Evaluators</td><td>GPT-4o-turbo</td><td>v1-2/25/2025</td><td>0 Run</td><td>May/13/2024 17:08</td><td></td></tr>
<tr><td>Log Evaluators</td><td>GPT-4o-turbo</td><td>v1-2/25/2025</td><td>3 Run(s)</td><td>May/13/2024 17:08</td><td></td></tr>
<tr><td>Accuracy Analyzer</td><td>GPT-4o-turbo</td><td>v3-2/25/2025</td><td>2 Run(s)</td><td>May/13/2024 17:0</td><td></td></tr>
<tr><td>Efficiency Assessor</td><td>GPT-4o-turbo</td><td>v3-2/25/2025</td><td>1 Run</td><td>May/13/2024 17:08</td><td></td></tr>
<tr><td>Test Evaluators</td><td>GPT-4o-mini</td><td>v2-2/25/2025</td><td>3 Run(s)</td><td>May/13/2024 17:08</td><td></td></tr>
<tr><td>Syntax Checker</td><td>GPT-4o-mini</td><td>v1-2/25/2025</td><td>1 Run</td><td>May/13/2024 17:08</td><td></td></tr>
</table>
</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</table>

2. Create a new version of an Evaluator FE-2354: [LLM Evaluator] Create a new Evaluator LAUNCHED

<table>
<tr>
<th>Requirement</th>
<th>Wireframe / Design</th>
</tr>
<tr>
<td>
Create a new Evaluator is a 2 steps form
1. Describe your Evaluator
2. <em>Mapping configuration (TBD)</em>
</td>
<td></td>
</tr>
<tr>
<td>
<strong>Describe</strong>
Attributes of an Evaluator includes:
1. Name
2. Prompt: write instruction to tune the LLM with guide on how to evaluate
3. Model: select base LLM model
4. Metric Config: Connect a Metric to be the core output of this evaluator
5. Reasoning: Add guide on justification of the evaluation result
</td>
<td>


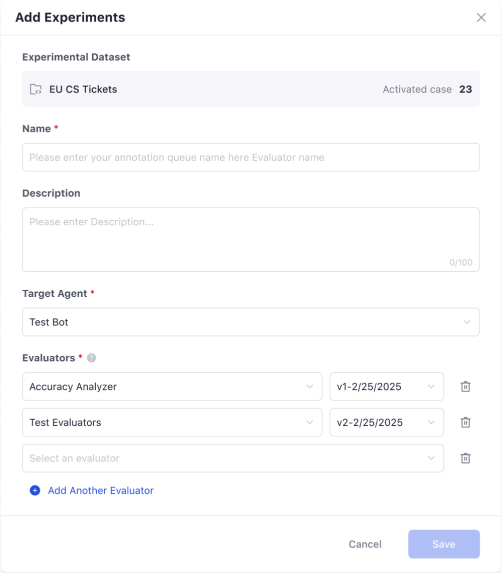
Describe your Evaluator
Evaluator Name *
Accurate analysis
Prompt *
Evaluate the correctness of the context on a continuous scale from 0 to 1. A context can be considered correct (Metric: 1) if it includes all the key facts from the ground truth and if every fact presented in the context is factually supported by the ground truth or common sense.
Example:
Query: Can eating carrots improve your vision?
Context: Everyone has heard, “Eat your carrots to have good eyesight!” Is there any truth to this statement or is it a bunch of baloney? Well no. Carrots won’t improve your visual acuity if you have less than perfect vision. A diet of carrots won’t give a blind person 20/20 vision. If your vision problems aren’t related to vitamin A, your vision won’t change no matter how many carrots you eat.
Ground truth: It depends. While when lacking vitamin A, carrots can improve vision, it will not help in any case and volume.
Metric: 0.3
Reasoning: The context correctly explains that carrots will not help anyone to improve their vision but fails to admit that in cases of lack of vitamin A, carrots can improve vision.
Input:
Query: {{query}}
Context: {{context}}
Ground truth: {{ground_truth}}
Think step by step.
The following variables are available:
query context ground_truth
Model *
GPT-4o-mini
Metric Config *
Accuracy
Only one option can be selected
Reasoning *
Provide a one sentence reasoning
</td>
</tr>
</table>

~~***Mapping***~~ ~~Allow user to set up variables that can be use in the evaluator instruction~~ ~~Support user map value from XXX to the variable~~ ~~*An example use case is that the user can set a variable Ground_Truth and expect the evaluator to extract that from the executed run to see if the answer is truth.*~~ ⚠ deprioritized	 Add Evaluator ✓ Describe your Evaluator > ② Mapping configuration Describe your Evaluator Mapping Reasoning * Evaluate the correctness of the context on a continuous scale from 0 to 1. A context can be considered correct (Metric: 1) if it includes all the key facts from the ground truth and if every fact presented in the context is factually supported by the ground truth or common sense. Example: Query: Can eating carrots improve your vision? Context: Everyone has heard, "Eat your carrots to have good eyesight!" Is there any truth to this statement or is it a bunch of baloney? Well no. Carrots won't improve your visual acuity if you have less than perfect vision. A diet of carrots won't give a blind person 20/20 vision. If your vision problems aren't related to vitamin A, your vision won't change no matter how many carrots you eat. Ground truth: It depends. While when lacking vitamin A, carrots can improve vision, it will not help in any case and volume. Metric: 0.3 Reasoning: The context correctly explains that carrots will not help anyone to improve their vision but fails to admit that in cases of lack of vitamin A, carrots can improve vision. Variable Mapping query Object *: Request Key *: query_calls JSON Path: query_calls[0].name query_calls[1].name Add JSON Path round_truth Object *: Request Key *: ground_truth_calls JSON Path: ground_truth_calls[0].name Add JSON Path ground_truth Object *: Request Key *: ground_truth_calls JSON Path: Please Input JSON Path Add JSON Path Previous Done
Leaving the prompt reset the content. Show confirmation pop up	 Close Configuration Content Page The current page configuration has not been saved, and the content that has been configured cannot be retrieved after closing this page. Cancel Confirm
User can create a new version of an existing evaluator	 PAPAGO.AI Texttexttexttextte... Construction: Bots, Evaluation (Human Annotation, LLM-as-a-judge, Dataset) Resources: Knowledge, Tools, Workflow Setting: Model Settings Charlotte Charlotte@gmail.com Eval Configs Metric Configs Evaluators Search... Name / Model New Evaluators / GPT-4o-turbo Log Evaluators / GPT-4o-turbo Accuracy Analyzer / GPT-4o-turbo Efficiency Assessor / GPT-4o-turbo Test Evaluators / GPT-4o-mini Syntax Checker / GPT-4o-mini Show 25 rows Accuracy Analyzer 2 Run(s) are currently using this template Model GPT-4o-mini Temperature: 0.80 Tokens Max: 4096 Metric Config Accuracy Correct,100% (0) Partially correct (1) Error,100% (2) provide a metric between 0 and 1 Reasoning provide a one sentence reasoning Prompt Evaluate the correctness of the context on a continuous scale from 0 to 1. A context can be considered correct (Metric: 1) if it includes all the key facts from the ground truth and if every fact presented in the context is factually supported by the ground truth or common sense. Example: Query: Can eating carrots improve your vision? Context: Everyone has heard, "Eat your carrots to have good eyesight!" Is there any truth to this statement or is it a bunch of baloney? Well no. Carrots won't improve your visual acuity if you have less than perfect vision. A diet of carrots won't give a blind person 20/20 vision. If your vision problems aren't related to vitamin A, your vision won't change no matter how many carrots you eat. Ground truth: It depends. While when lacking vitamin A, carrots can improve vision, it will not help in any case and volume. Metric: 0.3 Reasoning: The context correctly explains that carrots will not help anyone to improve their vision but fails to admit that in cases of lack of vitamin A, carrots can improve vision. Input: Query: {{query}} Context: {{context}} Ground truth: {{ground_truth}} v1-2/25/2025 + New Version

3. Add an Evaluator to an Experiment

A FE-2356: [Experiment] Add an Evaluator to an Experiment QA n LLM Evaluator will:

1. Have the goal to output a value for a attached Metric.
2. Follow the guide by user on how to measure the Metric.
3. It will review, analyze the input and output a value
4. It can also give justification on why such value is outputted.
5. The input are test Case Input, Expected Output and the Actual Output.

Requirement	Wireframe / Design
An evaluator can be add in a new experiment	 Add Experiments Experimental Dataset: EU CS Tickets — Activated case 23 Name *: Please enter your annotation queue name here Evaluator name Description: Please enter Description... 0/100 Target Agent *: Test Bot Evaluators *: Accuracy Analyzer — v1-2/25/2025 Test Evaluators — v2-2/25/2025 Select an evaluator Add Another Evaluator Cancel Save
An evaluator can be add in an executed experiment to give extra evaluation for such experiment	TBD

Adjustment

1. Mapping Test Case attribute in Evaluator prompt

User can reference `Experimenting Test Case’s Attributes` in the Evaluator prompts.

Experimenting Test Case’s Attributes

```
"experimentCase":{
  "actualOutput":"",
  "input":"",
  "expectedOutput":"",
  "context":"",
  "conversationHistory":"",
  "expectedToolCall":"",
}
```

Examples Evaluator Prompt

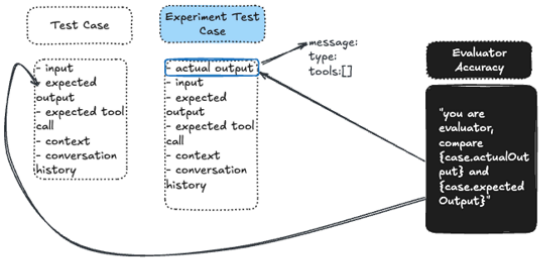
User Satisfaction Evaluator

```
You are a helpful AI bot that checks for user satisfaction based on the response text and its prompt. Here is the data:
```

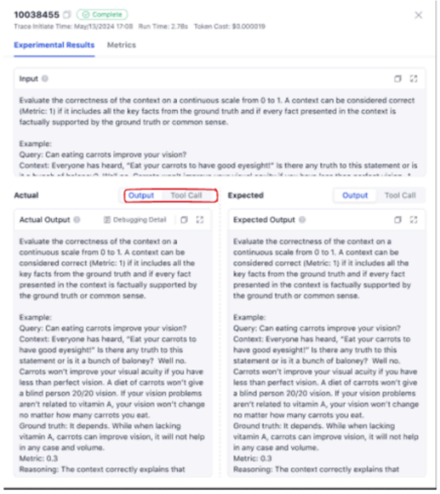
```
[BEGIN DATA]
============
[Response]: {actualOutput}
============
[Prompt]: {input}
[END DATA]
Compare the Response above to the Prompt and determine if the Response is satisfactory given the Prompt.
Your response must be a string, either SATISFIED or UNSATISFIED, and should not contain any text or characters aside from that.
The string UNSATISFIED means that the Response does not meet the user's needs or expectations based on the Prompt.
The string SATISFIED means the Response meets the user's needs or expectations based on the Prompt.

Then write out in a step by step manner an EXPLANATION to show how you determined if the user was satisfied or unsatisfied.
```

Accuracy Evaluator

```
You are a helpful AI bot that checks for Accuracy based on the Actual Output text and its Expected Output text. Here is the data:
[BEGIN DATA]
============
[Actual Output]: {testCase.actualOutput}
============
[Expected Output]: {testCase.expectedOutput}
[END DATA]
Compare the Actual Output above to the Expected Output and determine if the Actual Output is resembles the given Expected Output.
Your response must be a string, either LOW, Medium, and High, and should not contain any text or characters aside from that.
The string LOW means that the Actual Output does not resembling the content based on the Expected Output.
The string MEDIUM means that the Actual Output does resembling the content with minor vairation based on the Expected Output.
The string HIGH means that the Actual Output does resembling the content based on the Expected Output and replacible with that Expected Output content.

Then write out in a step by step manner an EXPLANATION to show how you determined if the Actual Output is LOW, MEDIUM, or HIGH.
```



"you are evaluator, compare {case.actualOutput} and {case.expectedOutput}"



10038455 Complete

Trace Initiate Time: May/13/2024 17:08 Run Time: 2.78s Token Cost: $0.000019

Input

Evaluate the correctness of the context on a continuous scale from 0 to 1. A context can be considered correct (Metric: 1) if it includes all the key facts from the ground truth and if every fact presented in the context is factually supported by the ground truth or common sense.

Example:
Query: Can eating carrots improve your vision?
Context: Everyone has heard, "Eat your carrots to have good eyesight!" Is there any truth to this statement or is

Actual

Actual Output

Evaluate the correctness of the context on a continuous scale from 0 to 1. A context can be considered correct (Metric: 1) if it includes all the key facts from the ground truth and if every fact presented in the context is factually supported by the ground truth or common sense.

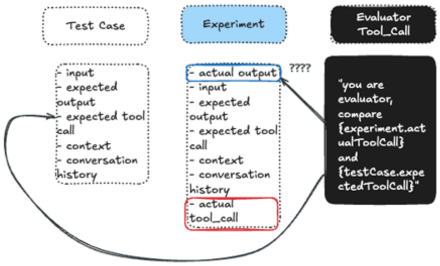
Example:
Query: Can eating carrots improve your vision?
Context: Everyone has heard, "Eat your carrots to have good eyesight!" Is there any truth to this statement or is it a bunch of baloney? Well no. Carrots won't improve your visual acuity if you have less than perfect vision. A diet of carrots won't give a blind person 20/20 vision. If your vision problems aren't related to vitamin A, your vision won't change no matter how many carrots you eat.
Ground truth: It depends. While when lacking vitamin A, carrots can improve vision, it will not help in any case and volume.
Metric: 0.3
Reasoning: The context correctly explains that

Expected

Expected Output

Evaluate the correctness of the context on a continuous scale from 0 to 1. A context can be considered correct (Metric: 1) if it includes all the key facts from the ground truth and if every fact presented in the context is factually supported by the ground truth or common sense.

Example:
Query: Can eating carrots improve your vision?
Context: Everyone has heard, "Eat your carrots to have good eyesight!" Is there any truth to this statement or is it a bunch of baloney? Well no. Carrots won't improve your visual acuity if you have less than perfect vision. A diet of carrots won't give a blind person 20/20 vision. If your vision problems aren't related to vitamin A, your vision won't change no matter how many carrots you eat.
Ground truth: It depends. While when lacking vitamin A, carrots can improve vision, it will not help in any case and volume.
Metric: 0.3
Reasoning: The context correctly explains that



"you are evaluator, compare {experiment.actualToolCall} and {testCase.expectedToolCall}"