[Evaluation] - Create Test Case and Manage with Datasets

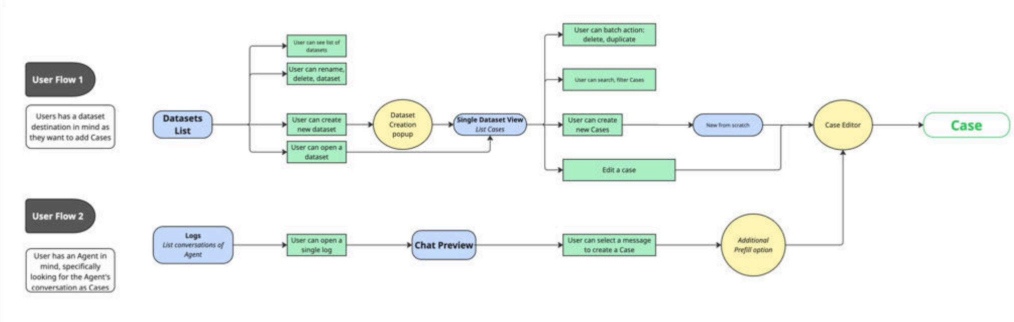
Research	Confident AI Blog - Resources to help teams stay confident in AI Blog - Langfuse Evaluating AI Agents - DeepLearning.AI
Wireframe	https://miro.com/app/board/uXjVIs9HNVg=/?moveToWidget=3458764633561380994&cot=14 Connect your Miro account
Design	Papago AI - Dataset
Relevant PRDs	[Evaluation] - 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. Evaluating or comparing agents requires running them through the same user **inputs** and then assessing the **outputs** against predefined metrics and criteria.

Solution

This new module will enable users to create **Datasets**, which are collections of **Cases** *(Test Cases).* Each Case allows users to record inputs, expected outputs, tool calls, and other relevant details as a single, consistent unit for future evaluation jobs.


User Flow 1
Users has a dataset destination in mind as they want to add Cases
Datasets List
User can see list of datasets
User can rename, delete, dataset
User can create new dataset
User can open a dataset
Dataset Creation popup
Single Dataset View
List Cases
User can batch action: delete, duplicate
User can search, filter Cases
User can create new Cases
Edit a case
New from scratch
Case Editor
Case
User Flow 2
User has an Agent in mind, specifically looking for the Agent's conversation as Cases
Logs
List conversations of Agent
User can open a single log
Chat Preview
User can select a message to create a Case
Additional Prefill option

MVP: User Flow Map - Datasets & Case Management

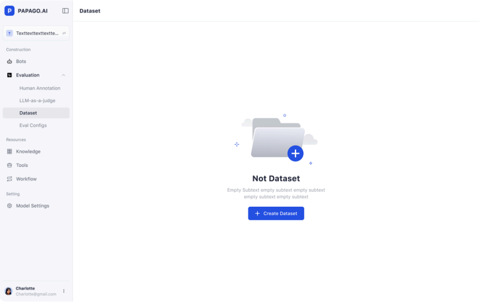
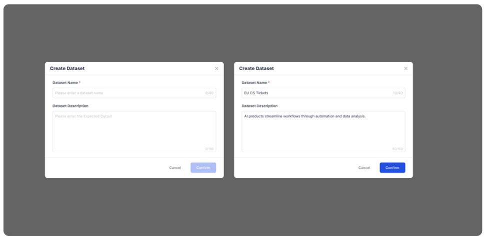
Experiment = **Dataset** + **Agent** + **Metrics (+ optional Evaluator).**

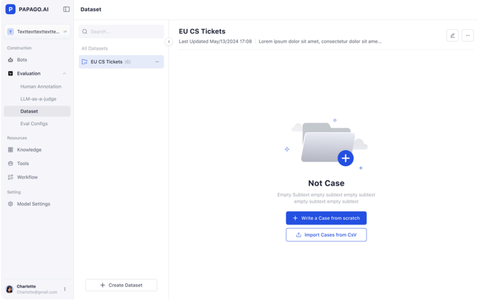
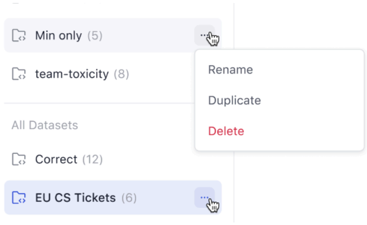
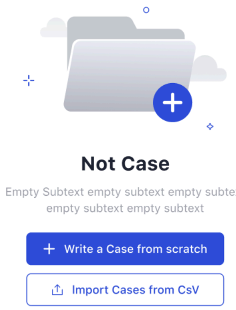
- Uses made-up *(totally synthetic, no live logs needed)* or hand-picked content from real log to create **test cases**.
- Lets users run the same **dataset of test cases** against **agents** to see who well it perform base on **Metrics**
- Results belong to the experiment itself, not the agent or any single message. Notice there's no "This message has been evaluated…" label here—everything rolls up to the experiment level.

We are building 3 related PRDs: Dataset, Metric and Experiment in this sprint

Requirement

1. Create a dataset FE-2340: [Case & Dataset] Create a dataset LAUNCHED

Requirement	Wireframe
Add Dataset as a new module on the platform, under Evaluation (also a new category). Empty state message: 1. Title: No Dataset 2. Sub-title: Dataset manage your test cases for agent evaluation.	 PAPAGO.AI Dataset Texttexttexttexte... Construction Bots Evaluation Human Annotation LLM-as-a-judge Dataset Eval Configs Resources Knowledge Tools Workflow Setting Model Settings Charlotte Charlotte@gmail.com + Create Dataset Not Dataset Empty Subtext empty subtext empty subtext empty subtext empty subtext
Create a new Dataset open the creation as pop up 1. Required Dataset Name 2. Data Description	 Create Dataset Dataset Name * Please enter a dataset name 0/40 Dataset Description Please enter the Expected Output 0/160 Cancel Confirm Create Dataset Dataset Name * 13/40 Dataset Description 60/160 Cancel Confirm EU CS Tickets AI products streamline workflows through automation and data analysis.

Datasets show a list of datasets as a row of times in a new left side pane. In Datasets view: 1. Each row is a specific dataset with a. name b. Number of cases per dataset 2. Datasets are searchable by name 3. The pane has two category: a. Recent Run b. All Dataset 4. User can Create a new dataset	 PAPAGO.AI Dataset Texttexttexttextte... Construction Bots Evaluation Human Annotation LLM-as-a-judge Dataset Eval Configs Resources Knowledge Tools Workflow Setting Model Settings Charlotte Charlotte@gmail.com Search... All Datasets EU CS Tickets (6) Create Dataset EU CS Tickets Not Case Write a Case from scratch Import Cases from CsV Last Updated May/13/2024 17:08 | Lorem ipsum dolor sit amet, consectetur dolor sit ame... Empty Subtext empty subtext empty subtext empty subtext empty subtext
Dataset can be Rename, Duplicate, and Delete	 Min only (5) team-toxicity (8) All Datasets Correct (12) EU CS Tickets (6) Rename Duplicate Delete
A new empty Dataset will prompt user to add cases 1. Write a Case from Scratch 2. Import Cases from CSV	 Not Case Write a Case from scratch Import Cases from CsV Empty Subtext empty subtext empty subtext empty subtext empty subtext

2. Cases in a dataset

User can see all Cases are listed in Cases tab.

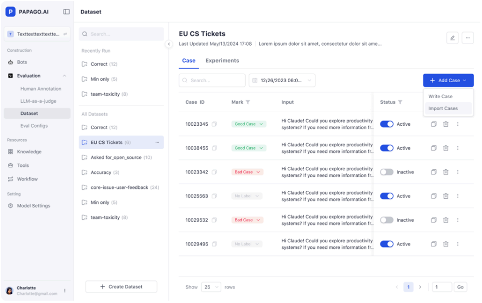
1. Sorted by latest edited on top
2. Number of columns equal to number of attributes
   a. Ensure Input Expected Output and Context is visible in the initial width
   b. Overflow is scrollable
3. Columns are case attributes:
   a. Case ID: Can be copy
   b. Mark: [Good,

      Bad] and **Filterable**
   c. Input, Expected Output and

      Context: Fix height, overflow hidden, expandable by click to snap to content length
   d. ~~Trace Initiate Time~~ Last Edited: date time edited **Sortable**
   e. Creator: Workspace member

      Filterable
   f. Status: Toggle [Active, InActive]

      Filterable
4. Freeze action column on the right to contain move, duplicate, and remove
   a. Move clone the case to another dataset
   b. Remove cases will have pop-up to confirm
   c. Duplicate will clone the selected case and added right

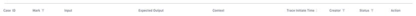


EU CS Tickets

Last Updated May/13/2024 17:08 | Lorem ipsum dolor sit amet, consectetur dolor sit ame...

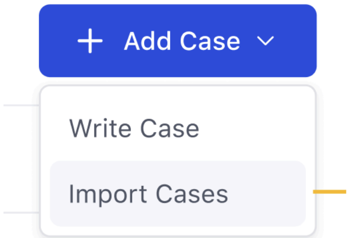
Case ID	Mark	Input	Status
10023345	Good Case	Hi Claude! Could you explore productivity systems? If you need more information fr...	Active
10038455	Good Case	Hi Claude! Could you explore productivity systems? If you need more information fr...	Active
10023342	Bad Case	Hi Claude! Could you explore productivity systems? If you need more information fr...	Inactive
10025563	No Label	Hi Claude! Could you explore productivity systems? If you need more information fr...	Active
10029532	Bad Case	Hi Claude! Could you explore productivity systems? If you need more information fr...	Inactive
10029495	No Label	Hi Claude! Could you explore productivity systems? If you need more information fr...	Active

Cases



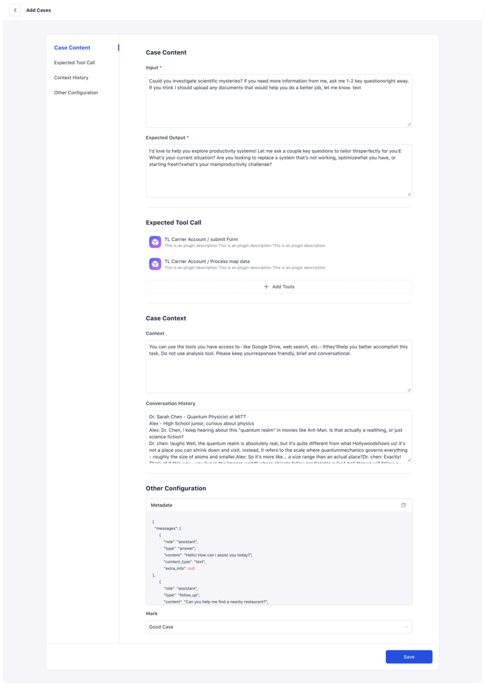
Case ID	Mark	Input	Expected Output	Context	Trace Initiate Time	Creator	Status	Action

Display columns

below. Re-sort the list only when user re-open the dataset 5. Cases view is configurable with: a. number of results to be show per page b. pagination between page 6. Cases can be search by name 7. Cases can be search	
Add a new case allow two option: 1. Write case (US3) 2. Import case (US4)	

3. Write case

Feature entry is US2’s `Add case` button

Requirement	Wireframe
Write case layout as a long form with section tree for quick navigation	

Add Cases

Case Content
Expected Tool Call
Context History
Other Configuration

Case Content

Input *

Could you investigate scientific mysteries? If you need more information from me, ask me 1-2 key questionsright away. If you think I should upload any documents that would help you do a better job, let me know. text

Expected Output *

I'd love to help you explore productivity systems! Let me ask a couple key questions to tailor thisperfectly for you:E What's your current situation? Are you looking to replace a system that's not working, optimizewhat you have, or starting fresh?xwhat's your mainproductivity challenae?

Expected Tool Call

TL Carrier Account / submit Form
This is an plugin description This is an plugin description This is an plugin description

TL Carrier Account / Process map data
This is an plugin description This is an plugin description This is an plugin description

+ Add Tools

Case Context

Context

You can use the tools you have access to- like Google Drive, web search, etc.- ifthey'llhelp you better accomplish this task, Do not use analysis tool. Please keep yourresponses friendly, brief and conversational.

Conversation History

Dr. Sarah Chen - Quantum Physicist at MITT
Alex - High School junior, curious about physics
Alex: Dr. Chen, I keep hearing about this "quantum realm" in movies like Ant-Man. Is that actually a realthing, or just science fiction?
Dr. chen: laughs Well, the quantum realm is absolutely real, but it's quite different from what Hollywoodshows us! it's not a place you can shrink down and visit. instead, it refers to the scale where quantummechanics governs everything - roughly the size of atoms and smaller.Alex: So it's more like... a size range than an actual place?Dr. chen: Exactly!

Other Configuration

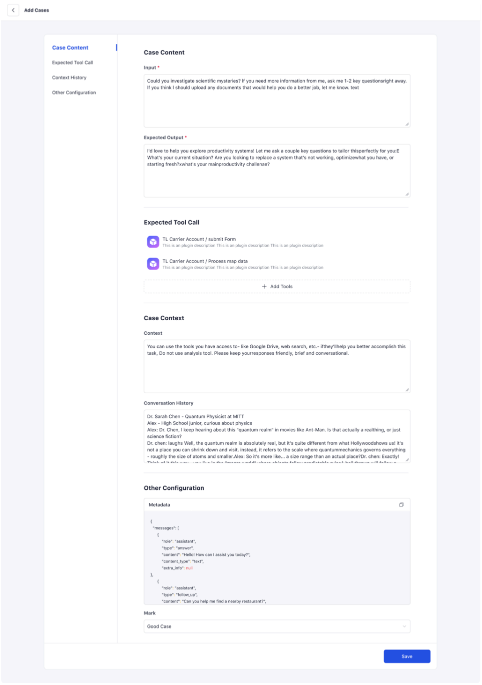
Metadata

```
{
 "messages": [
   {
     "role": "assistant",
     "type": "answer",
     "content": "Hello! How can I assist you today?",
     "content_type": "text",
     "extra_info": null
 },
   {
     "role": "assistant",
     "type": "follow_up",
     "content": "Can you help me find a nearby restaurant?",
```

Mark

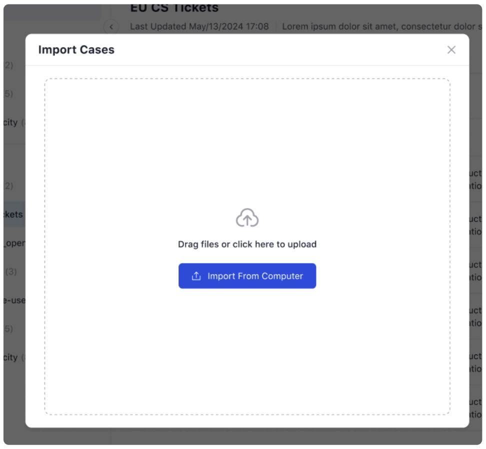
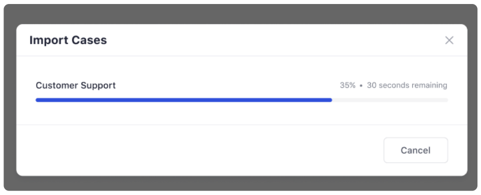
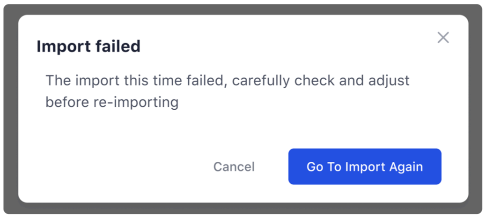
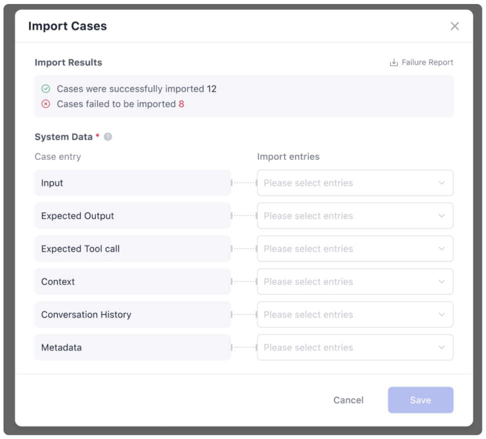
Good Case

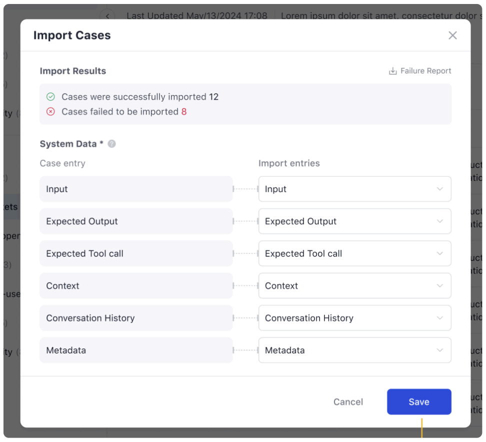
Save

User can fill the information of: 1. Input **- Required** 2. Expected output **- Required** 3. Expected tool call. (*Detail in next requirement)* 4. Context 5. Conversation History 6. Metadata (Optional) 7. Tags (Detail in next requirement) 8. Status (not visible) is set default to Active	 Add Cases Case Content Expected Tool Call Context History Other Configuration Case Content Input * Could you investigate scientific mysteries? If you need more information from me, ask me 1-2 key questionsright away. If you think I should upload any documents that would help you do a better job, let me know. text Expected Output * I'd love to help you explore productivity systems! Let me ask a couple key questions to tailor thisperfectly for you:E What's your current situation? Are you looking to replace a system that's not working, optimizewhat you have, or starting fresh?xwhat's your mainproductivity challenae? Expected Tool Call TL Carrier Account / submit Form This is an plugin description This is an plugin description This is an plugin description TL Carrier Account / Process map data This is an plugin description This is an plugin description This is an plugin description + Add Tools Case Context Context You can use the tools you have access to- like Google Drive, web search, etc.- ifthey'llhelp you better accomplish this task, Do not use analysis tool. Please keep yourresponses friendly, brief and conversational. Conversation History Dr. Sarah Chen - Quantum Physicist at MITT Alex - High School junior, curious about physics Alex: Dr. Chen, I keep hearing about this "quantum realm" in movies like Ant-Man. Is that actually a realthing, or just science fiction? Dr. chen: laughs Well, the quantum realm is absolutely real, but it's quite different from what Hollywoodshows us! it's not a place you can shrink down and visit. instead, it refers to the scale where quantummechanics governs everything - roughly the size of atoms and smaller.Alex: So it's more like... a size range than an actual place?Dr. chen: Exactly! Other Configuration Metadata { "messages": [{ "role": "assistant", "type": "answer", "content": "Hello! How can I assist you today?", "content_type": "text", "extra_info": null }, { "role": "assistant", "type": "follow_up", "content": "Can you help me find a nearby restaurant?", Mark Good Case Save
Each section must be collapsible due to the nature of lengthy content.	
Input box height is adjustable. This to make the text area length to adapt with content length.	
Input requirement to be validate on Save	
Each case has a system assigned unique Case ID	
Auto-save changes	
Content refresh when leave the page. Show confirmation that “Case has not been completed, do you want to leave”	
Save button will create a new case	

4. Import cases from CSV

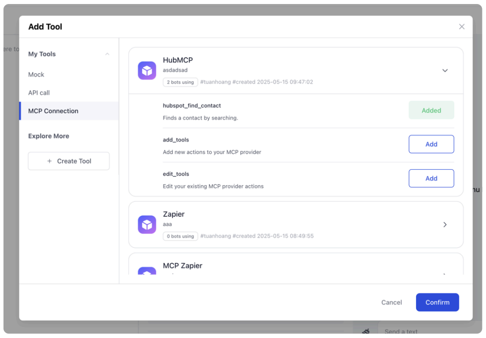
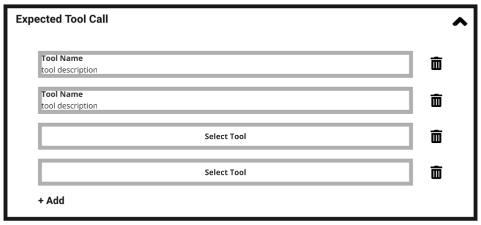
Requirement	**Wireframe**

User can upload cases from CSV. There is no restriction to the .csv file	 EU CS Tickets Last Updated May/13/2024 17:08 Lorem ipsum dolor sit amet, consectetur dolor s **Import Cases** Drag files or click here to upload Import From Computer
User will see upload status 1. Progressing: Percentage and Time Remaining 2. Failed: a. Show message b. Cancel and close to turn close the modal c. Go To Import Again will reopen the upload file widget	 **Import Cases** Customer Support 35% • 30 seconds remaining Cancel  **Import failed** The import this time failed, carefully check and adjust before re-importing Cancel Go To Import Again
On success, user can start mapping CSV column to case attributes. Partial success could happen, so showing how many case is success and not is also supported User can select any column in the uploaded file to map with the field in the case. The dropdown show a list of all .csv columns **Note to users:** Expected Tool Call is a string of multiple tool name, delimited by comma	 **Import Cases** Import Results Failure Report Cases were successfully imported 12 Cases failed to be imported 8 System Data * Case entry Import entries Input Please select entries Expected Output Please select entries Expected Tool call Please select entries Context Please select entries Conversation History Please select entries Metadata Please select entries Cancel Save

Examples: Input, Expected Output, Expected Tool Call xxx, qqqq, “tool_name_1, tool_name_2, tool_name_3” ccc, hh, tool_name_2	 Import Cases Import Results Failure Report Cases were successfully imported 12 Cases failed to be imported 8 System Data * Case entry: Input → Import entries: Input Case entry: Expected Output → Import entries: Expected Output Case entry: Expected Tool call → Import entries: Expected Tool call Case entry: Context → Import entries: Context Case entry: Conversation History → Import entries: Conversation History Case entry: Metadata → Import entries: Metadata Cancel Save

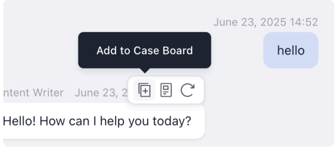
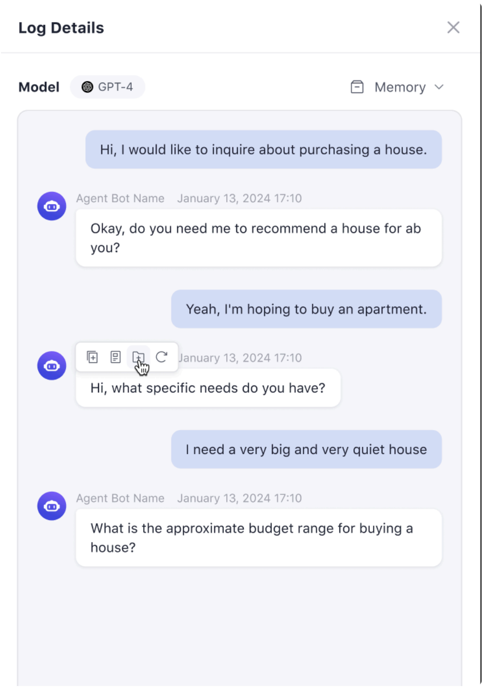
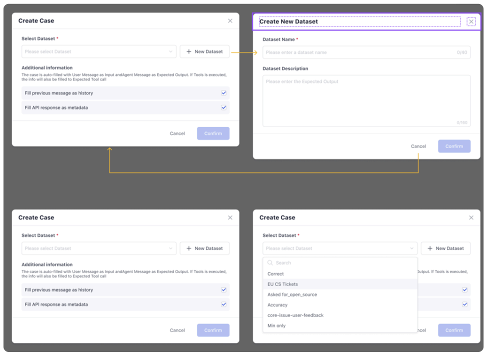
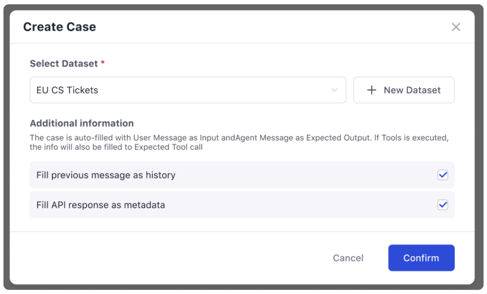
⚠ Adjustment: Not allow Metadata mapping win CSV import

5. Add tool calling to case

Requirement	**Wireframe**
Expected tool call section include add tool button. On click it open a list of tools that configured in the workspace *(Similar to existing selecting tools for agent, workflow).* On select, the tool name, description are captured as expected tool call (Copy by value not reference)	 Add Tool My Tools Mock API call MCP Connection Explore More + Create Tool HubMCP asdadsad 2 bots using #tuanhoang #created 2025-05-15 09:47:02 hubspot_find_contact Finds a contact by searching. Added add_tools Add new actions to your MCP provider Add edit_tools Edit your existing MCP provider actions Add Zapier aaa 0 bots using #tuanhoang #created 2025-05-15 08:49:55 MCP Zapier Cancel Confirm  Expected Tool Call Tool Name tool description Tool Name tool description Select Tool Select Tool + Add

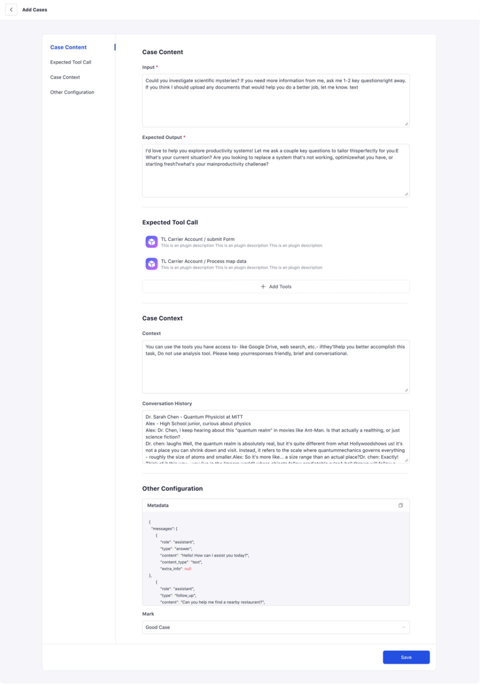
6. Create case from actual log

FE-2345: [Case & Dataset] Create case from actual log LAUNCHED **Requirement**	**Wireframe**

Change label `Add to Case Board` → `Copy clipboard`	 hello Hello! How can I help you today? June 23, 2025 14:52 Add to Case Board ntent Writer June 23, 2
In an Agent Detail Log, user can start adding case from a specific log chat preview panel Add Create new Case button to ALL message.	 **Log Details** **Model** GPT-4 Memory Hi, I would like to inquire about purchasing a house. Agent Bot Name January 13, 2024 17:10 Okay, do you need me to recommend a house for ab you? Yeah, I'm hoping to buy an apartment. January 13, 2024 17:10 Hi, what specific needs do you have? I need a very big and very quiet house Agent Bot Name January 13, 2024 17:10 What is the approximate budget range for buying a house?
`Create new case` open a Create Case editor 1. Select dataset Show list of dataset created a. Allow to Create new dataset. On dataset creation resume the case creation flo	 **Create Case** Select Dataset * Please select Dataset + New Dataset Additional information The case is auto-filled with User Message as Input andAgent Message as Expected Output. If Tools is executed, the info will also be filled to Expected Tool call Fill previous message as history ☑ Fill API response as metadata ☑ Cancel Confirm **Create New Dataset** Dataset Name * Please enter a dataset name 0/40 Dataset Description Please enter the Expected Output 0/160 Cancel Confirm **Create Case** Select Dataset * Please select Dataset + New Dataset Additional information The case is auto-filled with User Message as Input andAgent Message as Expected Output. If Tools is executed, the info will also be filled to Expected Tool call Fill previous message as history ☑ Fill API response as metadata ☑ Cancel Confirm **Create Case** Select Dataset * Please select Dataset + New Dataset Search Correct EU CS Tickets Asked for_open_source Accuracy core-issue-user-feedback Min only Cancel Confirm
User can select any message to add case. The platform will detect the relevant messages (User, Agent, Tool call) and filled in the new case. Specifically, the case is auto-filled with User Message as Input and Agent Message as Expected Output. If Tools is executed, the info will also be filled to Expected Tool call.	 **Create Case** Select Dataset * EU CS Tickets + New Dataset Additional information The case is auto-filled with User Message as Input andAgent Message as Expected Output. If Tools is executed, the info will also be filled to Expected Tool call Fill previous message as history ☑ Fill API response as metadata ☑ Cancel Confirm

User can also choice wether to add previous message as conversation history and fill the metadata with the whole API response from the server. Default value:

1. Fill history: On
    a. If user select message within the first conversation *(1 user message following by the agents responses)* history can be empty
2. Fill metadata: Off
    a. Meta on will copy the response JSON from our backend as metadata



Add Cases

Case Content

Expected Tool Call

Case Context

Other Configuration

Case Content

Input *

Could you investigate scientific mysteries? If you need more information from me, ask me 1-2 key questionsright away. If you think I should upload any documents that would help you do a better job, let me know. text

Expected Output *

I'd love to help you explore productivity systems! Let me ask a couple key questions to tailor thisperfectly for you:E What's your current situation? Are you looking to replace a system that's not working, optimizewhat you have, or starting fresh?xwhat's your mainproductivity challenae?

Expected Tool Call

TL Carrier Account / submit Form
This is an plugin description This is an plugin description This is an plugin description

TL Carrier Account / Process map data
This is an plugin description This is an plugin description This is an plugin description

+ Add Tools

Case Context

Context

You can use the tools you have access to- like Google Drive, web search, etc.- ifthey'llhelp you better accomplish this task, Do not use analysis tool. Please keep yourresponses friendly, brief and conversational.

Conversation History

Dr. Sarah Chen - Quantum Physicist at MITT
Alex - High School junior, curious about physics
Alex: Dr. Chen, I keep hearing about this "quantum realm" in movies like Ant-Man. Is that actually a realthing, or just science fiction?
Dr. chen: laughs Well, the quantum realm is absolutely real, but it's quite different from what Hollywoodshows us! it's not a place you can shrink down and visit. instead, it refers to the scale where quantummechanics governs everything - roughly the size of atoms and smaller.Alex: So it's more like... a size range than an actual place?Dr. chen: Exactly!

Other Configuration

Metadata

```
{
 "messages": [
   {
     "role": "assistant",
     "type": "answer",
     "content": "Hello! How can I assist you today?",
     "content_type": "text",
     "extra_info": null
 },
   {
     "role": "assistant",
     "type": "follow_up",
     "content": "Can you help me find a nearby restaurant?",
```

Mark

Good Case

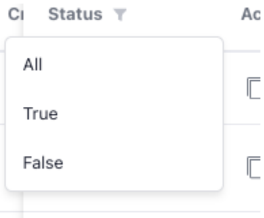
Save

Adjustment

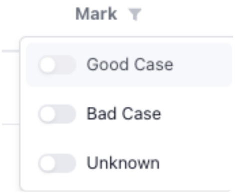
Filter UI element

Problem

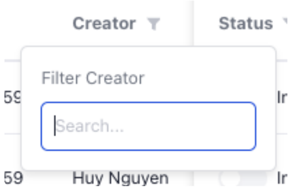
After testing on stage, I found that it is cumbersome to get to use to many UI style for the same filtering



❌ No indicator which status is currently filter



❌ Only one can be toggle on at a time.

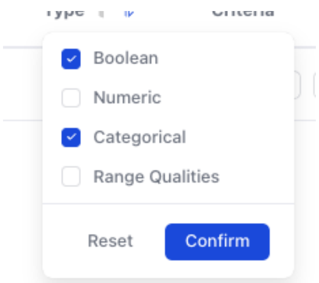


❌ List of all creators are known. Should show to select instead of search which is cumbersome and prone to typos

Suggestion

Re-use the filter in metrics that help:

1. Multiple filter apply
2. Clear indicator of which filter is applying
3. List of all possible filter criteria



☑ Boolean

□ Numeric

☑ Categorical

□ Range Qualities

Metric Filter