Get Started with LangSmith Multi-turn Evaluations

5 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=SC0KHJHJTP0
https://www.youtube.com/watch?v=sC0KhJHJTP0

SUMMARY

Langchain is launching a new feature called multi-turn evaluations, designed to assess user interactions over entire conversations rather than single messages. This feature enhances existing tools like evals and Linksmith, allowing for a comprehensive analysis of user intent, outcomes, and conversation trajectories.

- *Multi-turn evaluations are ideal for contexts requiring the understanding of entire user interactions, such as customer support chats.*
- *They allow for measuring user intent by clustering similar goals across conversations.*
- *Outcomes can be assessed, including user sentiment and whether their questions were satisfactorily answered.*
- *The trajectory of conversations can be analyzed to identify logical flow and potential issues, such as tool call loops.*
- *Configuration requires traces to be formatted correctly, including a list of messages for each turn.*
- *Users can filter traces based on conversation length to focus on more engaged interactions.*
- *Evaluators can be set up to analyze specific aspects, like user sentiment, with options to include or exclude certain message types.*
- *Multi-turn evaluators are now available for users to implement and provide feedback on.*

Hey, I'm Tanishi from Langchain. I'm excited to share a new feature that we're launching called multi-turn evaluations. These are used to run online evaluations over end to end user interactions. If you're already using evals and linksmith, these complement those and should be used when your evaluator needs the context of an entire thread or an entire user conversation rather than just the next message. So to walk through a quick example, let's say you have a chat app like a customer support agent. This is an example on the screen of a very short multi-turn conversation. You'll see that each turn is a human and AI pair. In Linksmith, one turn is represented by a trace and you can use our threads feature to group many traces into a single thread.

Multi-turn evaluators run over Linksmith threads. So what are the types of things that you can measure with multi-turn valves? There's broadly three buckets of things you can measure. First is clustering on intent. So trying to understand what the user is trying to accomplish. You can maybe break those out into common clusters and then measure those results over time. Second is understanding outcomes. This might be user sentiment across the entirety of a conversation or measuring if these were satisfied and their request was their request their question was answered. And then the third one is trajectory. This helps measure how a conversation unfolded or tool calls made in the logical order.

Were there you know tool call loops where the agent was stuck helps you measure things like that. Let's take a

look at what this looks like in practice. So I'm in the linksmith UI. I'm going to pull up a thread here. You can see this is an example of a trace that has four turns in the thread. And a couple things to call out with multi-turn evals. So before configuring multi-turn evals, one thing that we expect is your traces need to be in a specific format.

So specifically the top level input and output for each trace in a thread needs to have a list of messages. We do some nice things for you like automatically combine the messages if you're passing in a message history every time. And you'll see when we configure the evaluator that we have some nice pre-selected options that you can use in order to pass in specific parts of the message into the evaluator. The second thing you need when you're configuring thread level evals is to set an idle time for your tracing project.

And really what this means is what's the amount of time when a conversation or an interaction may be considered finished so that we can then send it to the evaluator and then let's go through actually configuring an evaluator. So if you go under the new evaluator tab, we have a new option here called evaluate a multi-thurn thread. I'm going to hit that. And in this example, I want to set up an evaluation for user sentiment across the thread.

So, one thing I'm going to do is I'm going to set a filter here. I'm going to filter out traces that have at least two turns. Reason being is we see a lot of single turn interactions in this app and those aren't as useful for measuring sentiment and typically the users that are most engaged will have longer conversations. So, let's focus on that. I'm going to create the prompt for my LLM as a judge evaluator. And once that's complete, you can move on to actually selecting which messages get sent into your evaluator. So we have three options. The first is all messages. So this would include all messages in your message array in both inputs and outputs. So things like tool calls will be included here. Human and AI pairs filters out things like tool calls or system prompts.

And then the last one we have here is first human and last AI message. For the sake of this evaluator, I only need human and AI pairs. So I'm going to select that. And then the last thing is to configure your feedback key. Just set it to user sentiment include reasoning. And I just want her to know was the user interaction positive or not. All right. And then we can create that evaluator. Now that that's configured, it'll show up in the evaluators tab here.

And things you can do when this evaluation runs on your threads, you can filter for threads. You might be wanting to look for threads that have a negative user sentiment because maybe your agent isn't working well in those cases. You can also create dashboards to track these results. Multi-turn evaluators are available for you to try out today. Give it a shot. Let us know what you think. There'll be some details in the description below on how to get started.