How to Run Multi-Modal Experiments in LangSmith Playground

5 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=-VONEPK3UWO
https://www.youtube.com/watch?v=-VOnEpk3uWo

SUMMARY

Katherine from Lang Chang demonstrates how to utilize Langmith Playground for multimodal data experiments, specifically focusing on extracting information from user receipts. The process involves creating a dataset with various receipt formats, defining prompt logic for extraction, and evaluating the performance of different models against predefined metrics.

- *Langmith Playground supports multimodal data, including text, images, PDFs, and audio.*
- *The demo focuses on extracting structured information from user receipts using a defined data schema.*
- *A dataset is created by uploading examples of receipts with input attachments and reference outputs.*
- *Prompt logic is pulled in to guide the language model (LM) in extracting relevant fields like name, date, merchant, and amount.*
- *Evaluation metrics are defined to assess the correctness of outputs against reference outputs, using a scoring system from 1 to 10.*
- *The experiment can be run with different models to determine which performs best with multimodal content.*
- *Results can be compared side by side, allowing for informed decisions on model selection and prompt iteration.*
- *The process emphasizes continuous improvement of quality metrics through iterative testing.*

Hi, this is Katherine from Lang Chang. Many real world use cases involve multimodal data. Not just text, but also images, PDFs, and audio. While most people use Langmith Playground to test and debug textbased agents, you can just as easily use it for multimodal agents as well. In this demo, we'll walk through running an experiment for multimodal prompt that extracts and process information from user receipt. To run an offline evaluation for multimodal content in Langmith, we'll first create a data set with input attachments and reference outputs. Next, we'll pull in the prompt logic that we want to test over. In this case, we'll be testing a prompt that ask the LM to identify information and extract based on predefined data schema.

Third, we need to define the metrics we want to evaluate performance. These are quality evaluators that reflect the key aspects of the output we care about. From there, we can view and compare experiment results from my evaluator runs. Now, let's hop over to LenMed to show the multimodal evaluation flow in the UI. As a first step, I can carry a multimodal data set by uploading in the data set view. Here I can create a new data set. We'll give it a name called multimodal receipt parsing. Hit create.

And here in the examples I can click on add examples where I add in the input reference output and add attachments to our data set. For our reference output we have listed a number of criterias that we would want our applications to parse. So that includes employee name, the date of receipt, merchant name, amount,

currency, the expense category and the description of the receipt. Hitting submit. This populates the example and I can preview the attachment by clicking on the image. Repeating this step. Now we have curated a data set with six examples with a mix of types of receipt in attachment formats from images to PDFs. For our second step, I've pulled in a prompt logic that I want to test in the prompt playground. Here we ask LM to extract structured output information and only ground the response based on the receipt. The fields here corresponds to the ideal output in the reference output that we have set up in our data set examples. And we have also enabled output schema to ensure the fields are outputed in the ideal format. This includes name, date, merchant, amount and currency, category, and description. With the data set and prompt to test on, we're ready to set up evaluation. Clicking on the set of evaluation button, this first prompts me to select our most recently created data set. And here scrolling to the bottom along with the human message, I can choose to select all the attachments from the data set. Next, we'll define our evaluator. For our use case, I will want to evaluate the correctness of our output against the reference output.

Linksmith provides a correctness evaluator out of the box that we can modify for our use case. First, rather than outputting a boolean, we can ask the LM to score on a scale of 1 to 10, ensuring that each field in a response is accurate, complete, and grounded. Towards the end of the prompt, we map our input and checks our generated output against the reference one. The prompt defaults the general input and output, but based on the context given on the right hand side, we can more specifically map the input to input question. In the bottom of the page, we'll also modify the schema, changing the boolean to a range of scores from 1 to 10. We're giving a definition that 10 refers to factually accurate and consistent results across all fields and one if all fields are missing information or have significant deviations from the reference output. From here, I can hit start and this will start populating first the generated outputs from a prompt based on the inputs and attachments. Once that's done, it will run the evaluators that then judges the outputs based on the reference outputs. Looks like our evaluators has finished running and we can view all scores in this playground view. And now specifically for a multimodal use case, I might want to test over different models to see which one interacts with multimodal content better. I can toggle to select a different model. In this case, let's try out enthropic. To rerun the set of experiment, I can hit start again.

And now that this has populated, I can directly click this link for full experiment. This will open up the experiment page for me in the data set view. I can compare side by side the reference outputs, the outputs as well as the overall correctness assessment. For any given trace, I can also click into the linksmith trace and in the trace view, we can also inspect the image that is being passed to the LLM. In addition to viewing a single experiment, I can go back to the summary experiment view to view summary stats across the two runs. And LinkSmith makes it easy to compare experiments side by side. I can inspect how the outputs differs and I can quickly toggle to see the four use cases that seems to have performed better using the ethropic model. And now having this information, I can make a better informed decision on the model and continue to iterate and test my prompt against the same data set to ensure that the quality metrics improve over time.