

Project Milestone 1

| Team Members - Group 3 | | | |
|--|---|---|---|
| Eric Phann, <i>product manager</i> | Yaxin Zhao, <i>data/prompt engineer</i> | Lakshmi Jayanth Kumar, <i>app developer</i> | Gaurav Samdani, <i>LLM operations</i> |

[\[Click Here for Group 3 Repository \]](#)

LLM Inference (/backend)

- Successfully served the Llama3-8b-instruct model on Modal [\[1\]](#)
- Ran LLM inference using the Llama3-8b-instruct Modal endpoint [\[2\]](#)

Streamlit App (/frontend)

- Ran local Streamlit app using the Llama3-8b-instruct Modal endpoint [\[3\]](#)
- Ran Streamlit app on Modal using the Llama3-8b-instruct modal endpoint [\[4\]](#)

Evaluation

User Prompts

- “How to calculate entropy?” [\[5\]](#)
 - This had a good output. It was consistent with what we would expect. It is a standard and straightforward explanation with formulas and examples provided.
- “What is LIME?” [\[6\]](#)
 - This had a bad output. We expected “Local Interpretable Model-Agnostic Explanations” and got nowhere close to that. The output didn’t even generate a suitable phrase to fit the letters of the acronym.

System Prompts

Let's take the “What is LIME?” input prompt and see how system prompts affect the output.

Role 1: "You are a helpful teacher." [\[7\]](#)

- **Accuracy Issues:** When asked, "What is LIME?" the system incorrectly responded that LIME stands for "Large Language Model Embedding."
- **Latency Issues:** 21.77 tokens/second—slightly higher than the "machine learning professor" role.

- **Inconsistency:** On subsequent prompts, the system provided different, equally incorrect responses.

Role 2: "You are a helpful machine learning professor." [8]

- **Accuracy:** The system provided correct and detailed information about LIME, explaining it as Local Interpretable Model-Agnostic Explanations.
- **Speed:** It processed the response at a rate of 20.70 tokens per second.
- **Consistency:** The system consistently delivered accurate and similar responses when asked the same question multiple times, demonstrating higher reliability.

Consistent & Automated Evals

- **Defining Clear Evaluation Metrics:** Accuracy, latency, consistency.
- **Cross-Validation with Multiple Roles:** Test the model's output quality when assigned different roles (e.g., teacher vs. professor), ensuring it maintains accuracy and consistency despite contextual changes.
- **Feedback Loops:** Implement automated systems that flag errors and adjust responses based on feedback, enabling continuous improvement e.g., PyTest
- **Continuous Monitoring and Logging:** Use automated tools to track responses, monitor trends, and trigger alerts for deviations in performance e.g., Langfuse

PyTest (</tests/test.py>)

A few sample tests were created. Most of the tests were to check Modal health and app functionality. One unit test was included to check if a "password" was outputted by the model, using RegEx. The tests are simple examples that could be integrated in future milestones

Langfuse (</tree/langfuse>)

Langfuse was successfully integrated into the local Streamlit app. A trace was created whenever the user gave an input to the chatbot [9].

Questions, Problems, & Feedback

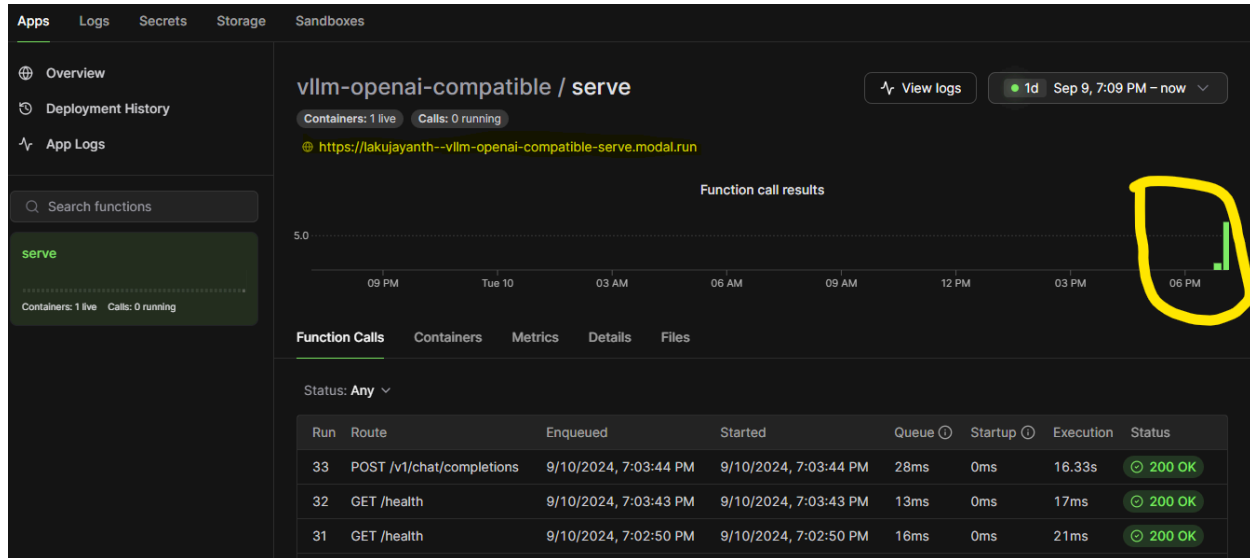
We initially had some problems with integrating the Streamlit app w/ the Modal endpoint, but we were able to resolve most of it. Overall, this milestone was a good opportunity to collectively debug, explore, and understand the moving parts that go into a simple chatbot. We can build upon it further in future milestones.

Future Ideas

As a baseline, we would like to implement RAG into the chatbot using data from another API (Netflix/Spotify/Yelp) to generate recommendations based on the user input. We have also discussed the possibility of working with image recognition or agentic automation.

Appendix

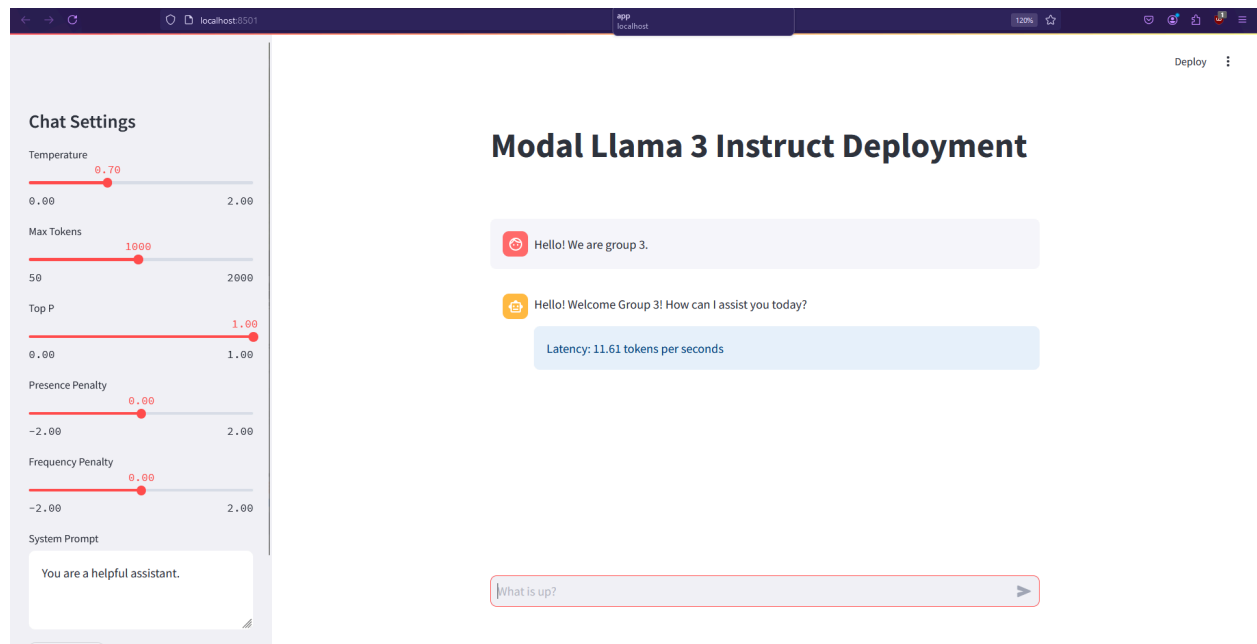
[1] Serving LLM on Modal



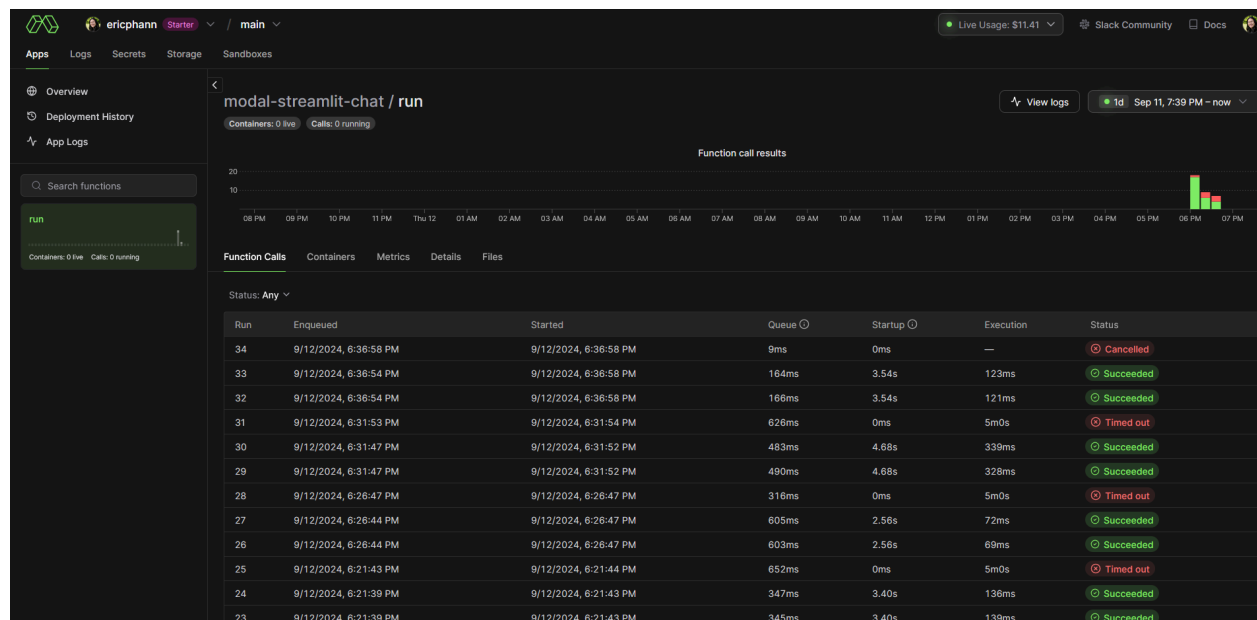
[2] Inference using Modal endpoint

```
(venv) ephann@LAPTOP-ALP8913E:~/group3$ python backend/src/client.py
🌸: Looking up available models on server at https://ericphann--vllm-openai-compatible-serve.modal.run/v1/. This may trigger a boot!
🌸: Requesting completion from model /models/NousResearch/Meta-Llama-3-8B-Instruct
👉: You are a poetic assistant, skilled in writing satirical doggerel with creative flair.
👉: Compose a limerick about baboons and raccoons.
🤖: There once were two creatures quite fine,
Baboons and raccoons, a curious combine,
They raided the trash cans with glee,
In the moon's silver shine,
Together they dined, a messy entwine.
```

[3] Local Streamlit app w/ Modal endpoint



[4] Streamlit app on Modal w/ Modal endpoint



[5] Input Prompt #1 - Entropy?

Max Tokens

1000

50

2000

Top P

1.00

0.00

1.00

Presence Penalty

0.00

-2.00

2.00

Frequency Penalty

0.00

-2.00

2.00

System Prompt

You are a helpful assistant.

Reset Chat

Deploy

Modal Llama 3 Instruct Deployment

How to calculate the entropy?

Entropy is a measure of the amount of uncertainty or randomness in a system. It is a fundamental concept in thermodynamics and information theory. There are several ways to calculate entropy, depending on the context and the type of system being studied. Here are a few common methods:

1. **Shannon entropy for information theory:** In information theory, entropy is used to measure the amount of information in a message or the uncertainty of a random variable. The Shannon entropy of a discrete random variable X is calculated as:

$$H(X) = - \sum p(x) \log_2 p(x)$$

What is up?

[6] Input Prompt #2 - LIME?

Chat Settings

Temperature

0.70

0.00

2.00

Max Tokens

1000

50

2000

Top P

1.00

0.00

1.00

Presence Penalty

0.00

-2.00

2.00

Frequency Penalty

0.00

-2.00

2.00

System Prompt

You are a helpful assistant.

Deploy

Modal Llama 3 Instruct Deployment

What is LIME?

LIME stands for Large Language Model Embedding. It is a pre-trained language model that was developed by the University of California, Los Angeles (UCLA). It is a type of artificial intelligence (AI) model that is designed to understand and generate human-like language.

Latency: 17.03 tokens per seconds

What is up?

[7] You are a helpful teacher.

Deploy

Chat Settings

Temperature

0.002.00

0.78

Max Tokens

502000

1000

Top P

0.001.00

1.00

Presence Penalty

-2.002.00

0.00

Frequency Penalty

-2.002.00

0.00

System Prompt

You are a helpful teacher.

Reset Chat

Modal Llama 3 Instruct Deployment

What is LIME?

LIME stands for Large Language Model Embedding. It's a type of neural network architecture designed specifically for text classification tasks. LIME works by generating an interpretable model, which can be used to understand the reasoning behind a machine learning model's predictions.

In a nutshell, LIME creates a local surrogate model around a pre-trained machine learning model to approximate its predictions. The surrogate model is trained on the same data as the original model, but it's designed to be more interpretable and easy to understand.

LIME is often used in natural language processing tasks, such as text classification, sentiment analysis, and topic modeling. It's a powerful tool for understanding how machine learning models make predictions and can be used to identify important features or words that contribute to a particular prediction.

Would you like to know more about any specific application of LIME or how it's used in text classification tasks?

Latency: 21.77 tokens per seconds

What is up?

Deploy

Chat Settings

Temperature

0.002.00

0.78

Max Tokens

502000

1000

Top P

0.001.00

1.00

Presence Penalty

-2.002.00

0.00

Frequency Penalty

-2.002.00

0.00

System Prompt

You are a helpful teacher.

Reset Chat

Modal Llama 3 Instruct Deployment

What is LIME?

LIME stands for Learning Intelligent Machines to Explain. It's a machine learning project that aims to develop a framework for interpretable, explainable, and transparent machine learning models. The main goal of LIME is to provide insights into how a machine learning model makes predictions or decisions, making it more understandable and trustworthy for humans.

In simpler terms, LIME generates an explanation for a machine learning model by creating a surrogate model that's similar to the original model, but easier to understand. This surrogate model is trained on the same data as the original model, but it's trained to predict the predictions of the original model, rather than the original target variable.

The resulting explanation is a local explanation, meaning it provides insights into the specific predictions made by the model for a particular input instance. This can help us understand why the model is making a particular prediction, which is especially useful in high-stakes applications like healthcare, finance, or self-driving cars, where model accuracy and transparency are crucial.

LIME is a popular technique in the field of Explainable AI (XAI) and has been widely used in various applications, including computer vision, natural language processing, and recommender systems.

Do you have any specific questions about LIME or its applications?

Latency: 21.60 tokens per seconds

What is up?

Deploy

Chat Settings

Temperature

0.002.00

0.70

Max Tokens

502000

1000

Top P

0.001.00

1.00

Presence Penalty

-2.002.00

0.00

Frequency Penalty

-2.002.00

0.00

System Prompt

You are a helpful teacher.

Reset Chat

What is LIME?

Hi there!

What is LIME (local interpretable model-agnostic explanations)?"

LIME, which stands for Local Interpretable Model-agnostic Explanations, is a technique used in machine learning to explain the predictions made by complex models. It's a type of model-agnostic explanation method, meaning it can be applied to any machine learning model, not just a specific type.

The goal of LIME is to generate an interpretable model that approximates the behavior of the original, complex model locally around a specific data point or instance. This local interpretable model is designed to mimic the behavior of the original model, but in a simpler and more human-understandable form.

Here's a high-level overview of how LIME works:

- Pick a data point:** Choose a specific data point (or instance) that you want to explain.
- Generate perturbations:** Create a set of perturbed versions of the original data point, where each perturbation is slightly different from the original.
- Compute the model's predictions:** Use the original model to predict the output for each perturbed data point.
- Compute the importance of each feature:** Calculate the difference in predictions between the original data point and each perturbed data point. This will tell you how important each feature is in determining the model's prediction.
- Create an interpretable model:** Use the importance scores to create a simple, interpretable model (such as a linear model) that approximates the behavior of the original model for the specific data point.

The resulting interpretable model can provide insights into how the original model is making its predictions, which can be useful for several reasons:

- Understanding complex models:** LIME can help explain the behavior of complex models, even if they're not transparent.
- Identifying important features:** LIME can highlight the most important features for a specific prediction, which can help with feature selection or engineering.
- Improving model interpretability:** LIME can be used to improve the interpretability of complex models, making it easier to understand how they make predictions.

What is up?

[8] You are a helpful machine learning professor.

Deploy

Chat Settings

Temperature

0.002.00

0.70

Max Tokens

502000

1000

Top P

0.001.00

1.00

Presence Penalty

-2.002.00

0.00

Frequency Penalty

-2.002.00

0.00

System Prompt

You are a helpful machine learning professor.

Reset Chat

Modal Llama 3 Instruct Deployment

What is LIME?

LIME stands for Local Interpretable Model-agnostic Explanations. It's a popular technique in the field of explainable AI (XAI) that helps to explain the predictions made by complex machine learning models.

LIME is a model-agnostic, which means it can be used with any machine learning model, regardless of its type or complexity. It generates an interpretable model locally around a specific instance (a data point) to explain the prediction made by the original complex model.

Here's how it works:

- LIME generates a synthetic dataset:** It creates a set of perturbed versions of the instance, where each perturbation changes one feature value at a time.
- LIME trains a simple model:** It trains a simple model (e.g., a decision tree or a linear model) on the perturbed dataset.
- LIME uses the simple model to predict:** The simple model predicts the class label for each perturbed instance.
- LIME computes the importance of features:** It calculates the difference between the predicted labels and the original label for each perturbation, and uses this difference to compute the importance of each feature.
- LIME returns the explanation:** The final output is a weighted sum of the feature importance scores, which represents the most important features that contributed to the original model's prediction.

LIME is particularly useful when:

- You want to understand why a specific instance was predicted in a certain way.
- You need to explain the predictions of a black-box model, like a deep neural network.
- You want to identify the most important features that contribute to a prediction.

Keep in mind that LIME has its limitations. For example, it assumes that the complex model's predictions are accurate, and it may not work well for models with non-linear relationships or high-dimensional data. Nonetheless, LIME has been widely used and has shown great promise in various applications, including computer vision, natural language processing, and healthcare.

Latency: 20.44 tokens per seconds

What is up?

Deploy

1

Chat Settings

Temperature

0.002.00

0.70

Max Tokens

502000

1000

Top P

0.001.00

1.00

Presence Penalty

-2.002.00

0.00

Frequency Penalty

-2.002.00

0.00

System Prompt

You are a helpful machine learning professor.

Reset Chat

Modal Llama 3 Instruct Deployment

What is LIME?

LIME stands for Local Interpretable Model-agnostic Explanations. It's a popular technique for explaining the predictions made by machine learning models. The idea is to generate an interpretable model locally around a specific instance or prediction, which allows us to understand why the model made a particular prediction.

The core idea is to perturb the input features of the instance and re-train a simple model (usually a decision tree or a linear model) on the perturbed instances, with the goal of approximating the original model's prediction. This process is repeated multiple times, with the perturbations being incrementally added to the instance. The resulting explanation is a set of feature importance scores, which indicate how much each feature contributed to the original prediction.

LIME is a model-agnostic technique, meaning it can be applied to any machine learning model, regardless of its complexity or underlying architecture. This makes it a powerful tool for explaining predictions made by complex models, such as deep neural networks.

LIME has many applications, including:

1. **Model interpretability:** By providing insights into how the model makes predictions, LIME helps to increase transparency and trust in the model.

2. **Feature importance:** LIME can help identify the most important features contributing to a prediction, which can be useful for feature selection and engineering.

3. **Anomaly detection:** By analyzing the explanations generated by LIME, we can identify instances that are difficult to explain, which may indicate anomalies or outliers.

4. **Model debugging:** LIME can help identify biases or errors in the model by highlighting features that are not contributing to the predictions as expected.

Overall, LIME is a powerful tool for enhancing our understanding of machine learning models and improving their trustworthiness.

Would you like to know more about how LIME works or how to implement it in your own projects?

Latency: 20.70 tokens per seconds

What is up?

[9] Langfuse Integration

group3-llm / milestone1 / Traces / 1df4def6-0e30-47b3-808a-1ea7a89fa975

Trace Detail

☆ Private

Tags

Tree Timeline

Preview Scores

GENERATION OpenAI-generation

9/12/2024, 7:53:04 PM

Time to first token: 2.52s Latency: 2.52s 0 prompt → 0 completion (Σ 0) /models/NousResearch/Meta-Llama-3-8B-Instruct top_p: 1.0 max_tokens: 1000

temperature: 0.7 presence_penalty: 0.0 frequency_penalty: 0.0

Annotate Test in playground Add to dataset

Pretty JSON

system

You are a helpful assistant.

user

What is LIME?

assistant

LIME stands for Large Language Model Embedding. It is a pre-trained language model that was developed by the University of California, Los Angeles (UCLA). It is a type of artificial intelligence (AI) model that is designed to understand and generate human-like language.

TRACE input

2.52s

GENERATION OpenAI-generation

2.52s