Model evaluation metrics

LLM Evaluation - Challenges

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

LLM Evaluation - Challenges

"Mike really loves drinking tea."

❤️☕

=

"Mike adores sipping tea."

❤️☕

"Mike does not drink coffee."

🤢☕

≠

"Mike does drink coffee."

🧐☕

LLM Evaluation - Metrics

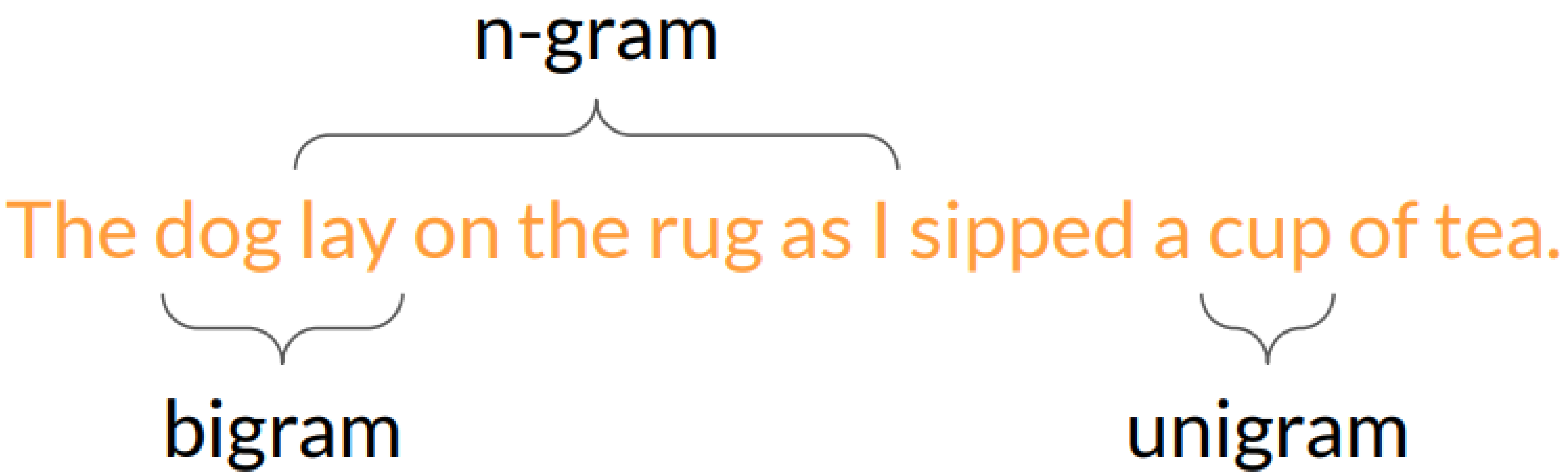
ROUGE

- Used for text summarization
- Compares a summary to one or more reference summaries

BLEU SCORE

- Used for text translation
- Compares to human-generated translations

LLM Evaluation - Metrics - Terminology



n-gram

The dog lay on the rug as I sipped a cup of tea.

bigram

unigram

LLM Evaluation - Metrics - ROUGE-1

Reference (human):

It is cold outside.

Generated output:

It is not cold outside.

$$\text{ROUGE-1 Recall} = \frac{\text{unigram matches}}{\text{unigrams in reference}} = \frac{4}{4} = 1.0$$

$$\text{ROUGE-1 Precision:} = \frac{\text{unigram matches}}{\text{unigrams in output}} = \frac{4}{5} = 0.8$$

$$\text{ROUGE-1 F1:} = 2\,\frac{\text{precision x recall}}{\text{precision + recall}} = 2\,\frac{0.8}{1.8} = 0.89$$

LLM Evaluation - Metrics - ROUGE-2

Reference (human):

It is cold outside.

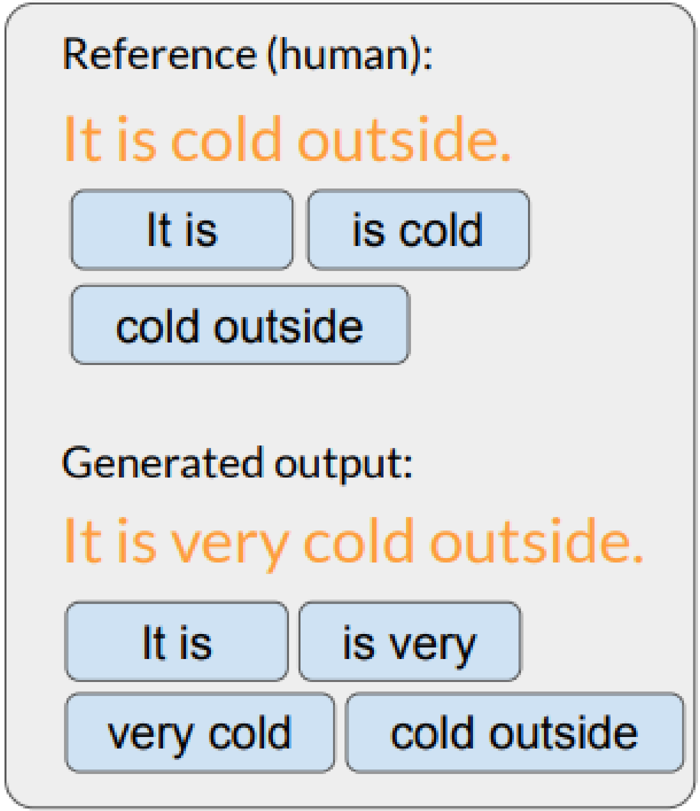
It is | is cold | cold outside

Generated output:

It is very cold outside.

It is | is very | very cold | cold outside

LLM Evaluation - Metrics - ROUGE-2



Reference (human):

It is cold outside.

It is

is cold

cold outside

Generated output:

It is very cold outside.

It is

is very

very cold

cold outside

$$\text{ROUGE-2 Recall:} = \frac{\text{bigram matches}}{\text{bigrams in reference}} = \frac{2}{3} = 0.67$$

$$\text{ROUGE-2 Precision:} = \frac{\text{bigram matches}}{\text{bigrams in output}} = \frac{2}{4} = 0.5$$

$$\text{ROUGE-2 F1:} = 2\,\frac{\text{precision x recall}}{\text{precision + recall}} = 2\,\frac{0.335}{1.17} = 0.57$$

LLM Evaluation - Metrics - ROUGE-L

Reference (human):

It is cold outside.

Generated output:

It is very cold outside.

Longest common subsequence (LCS):

It is

cold outside

2

LLM Evaluation - Metrics - ROUGE-L

Reference (human):

It is cold outside.

Generated output:

It is very cold outside.

LCS:
Longest common subsequence

$$\text{ROUGE-L Recall:} = \frac{\text{LCS(Gen, Ref)}}{\text{unigrams in reference}} = \frac{2}{4} = 0.5$$

$$\text{ROUGE-L Precision:} = \frac{\text{LCS(Gen, Ref)}}{\text{unigrams in output}} = \frac{2}{5} = 0.4$$

$$\text{ROUGE-L F1:} = 2\,\frac{\text{precision x recall}}{\text{precision + recall}} = 2\,\frac{0.2}{0.9} = 0.44$$

LLM Evaluation - Metrics

ROUGE

- Used for text summarization
- Compares a summary to one or more reference summaries

BLEU SCORE

- Used for text translation
- Compares to human-generated translations

LLM Evaluation - Metrics - BLEU

BLEU metric = Avg(precision across range of n-gram sizes)

Reference (human):

I am very happy to say that I am drinking a warm cup of tea.

Generated output:

I am very happy that I am drinking a cup of tea. - BLEU 0.495

I am very happy that I am drinking a warm cup of tea. - BLEU 0.730

I am very happy to say that I am drinking a warm tea. - BLEU 0.798

I am very happy to say that I am drinking a warm cup of tea. - BLEU 1.000

Benchmarks

Evaluation benchmarks

GLUE

SuperGLUE

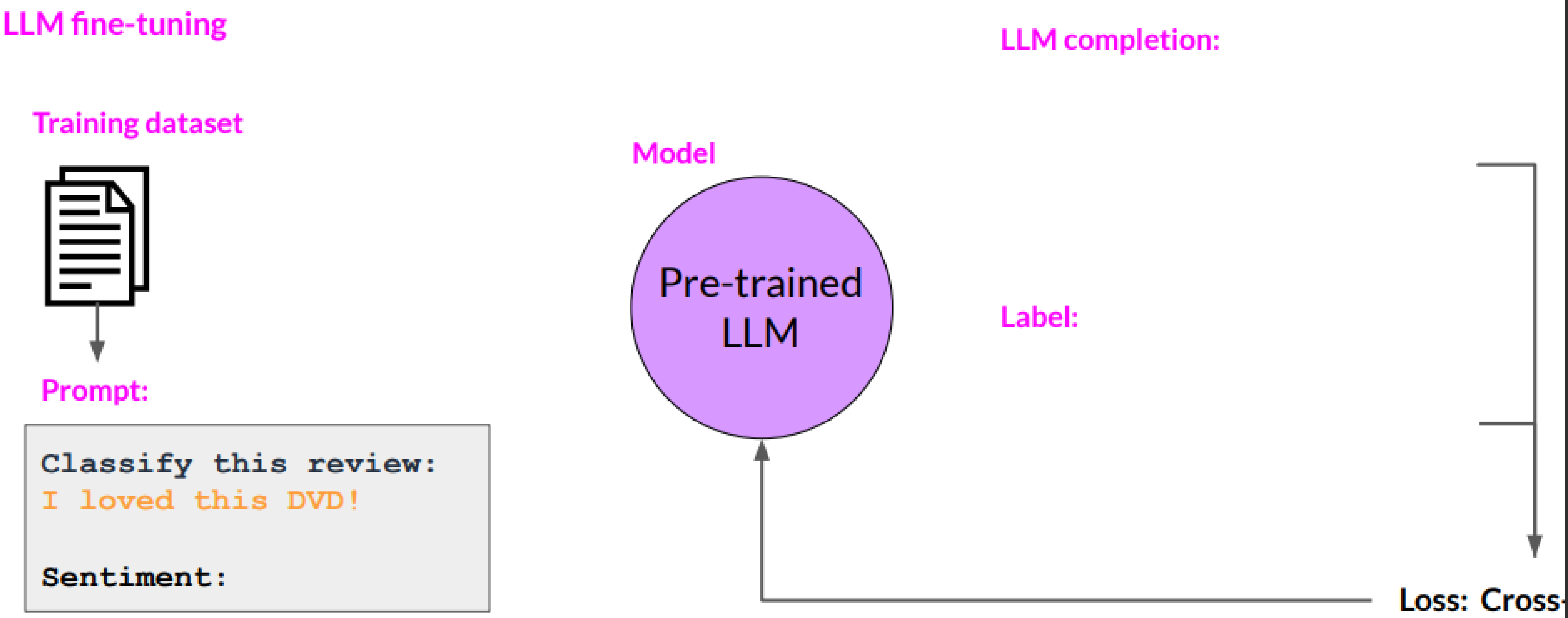
HELM

MMLU (Massive Multitask Language Understanding)

BIG-bench 🪑

Key takeaways

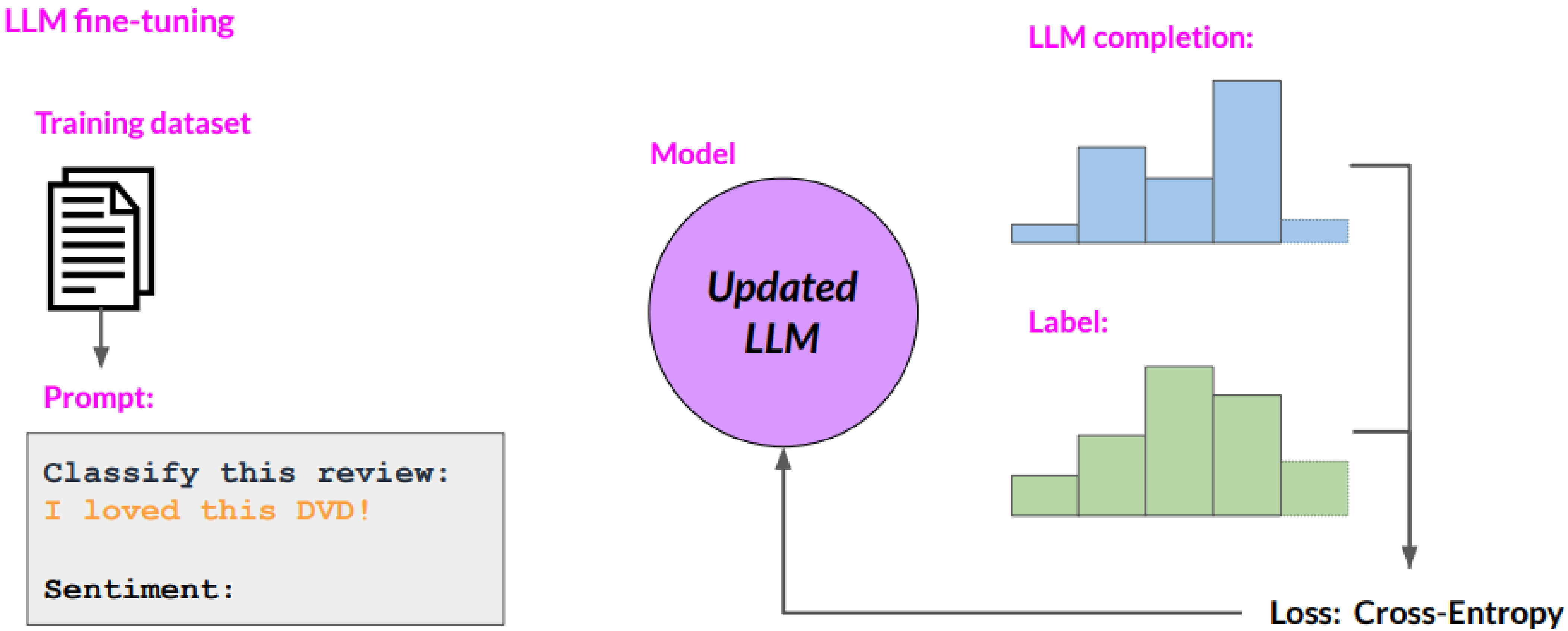
LLM fine-tuning process



```
Classify this review:
I loved this DVD!

Sentiment:
```

LLM fine-tuning process



LLM fine-tuning

Training dataset

Prompt:

```
Classify this review:
I loved this DVD!

Sentiment:
```

Model

Updated LLM

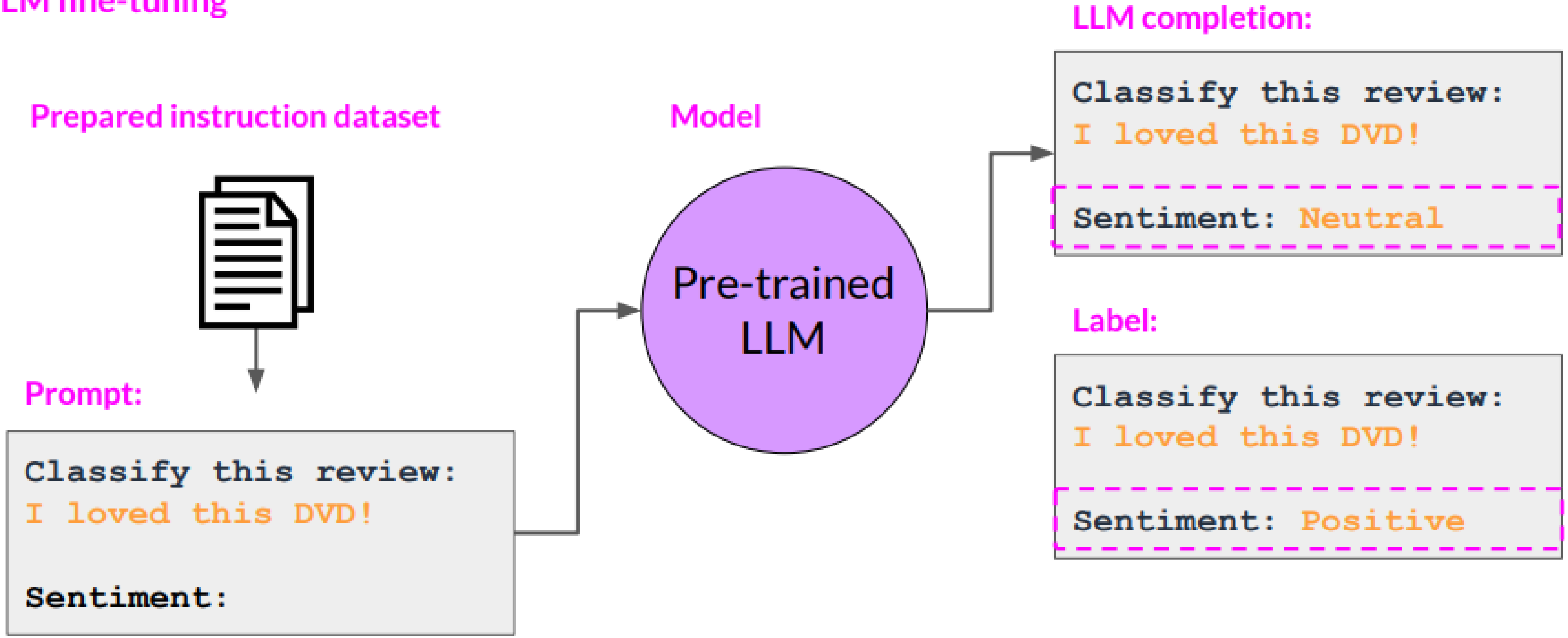
LLM completion:

Label:

Loss: Cross-Entropy

LLM fine-tuning process

LLM fine-tuning



Prepared instruction dataset

Prompt:

```
Classify this review:
I loved this DVD!

Sentiment:
```

Model

Pre-trained LLM

LLM completion:

```
Classify this review:
I loved this DVD!

Sentiment: Neutral
```

Label:

```
Classify this review:
I loved this DVD!

Sentiment: Positive
```