Automating Prompt Engineering for Forecasting Tasks

Coco Cao, Severin Perez, Luc Robitaille, Greg Rosen
DATASCI 210 - Capstone, Section 3, Summer 2023
U.C. Berkeley, Master of Information and Data Science Program

Berkeley
UNIVERSITY OF CALIFORNIA

Team

Coco Cao

Severin Perez

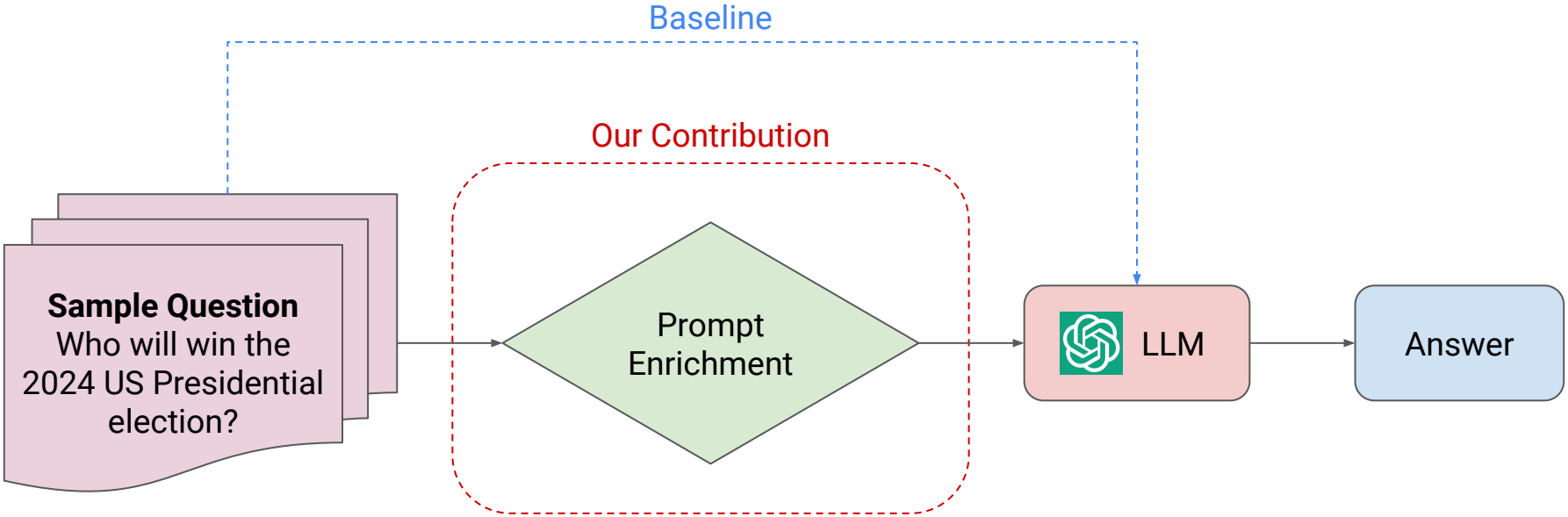
Luc Robitaille

Greg Rosen

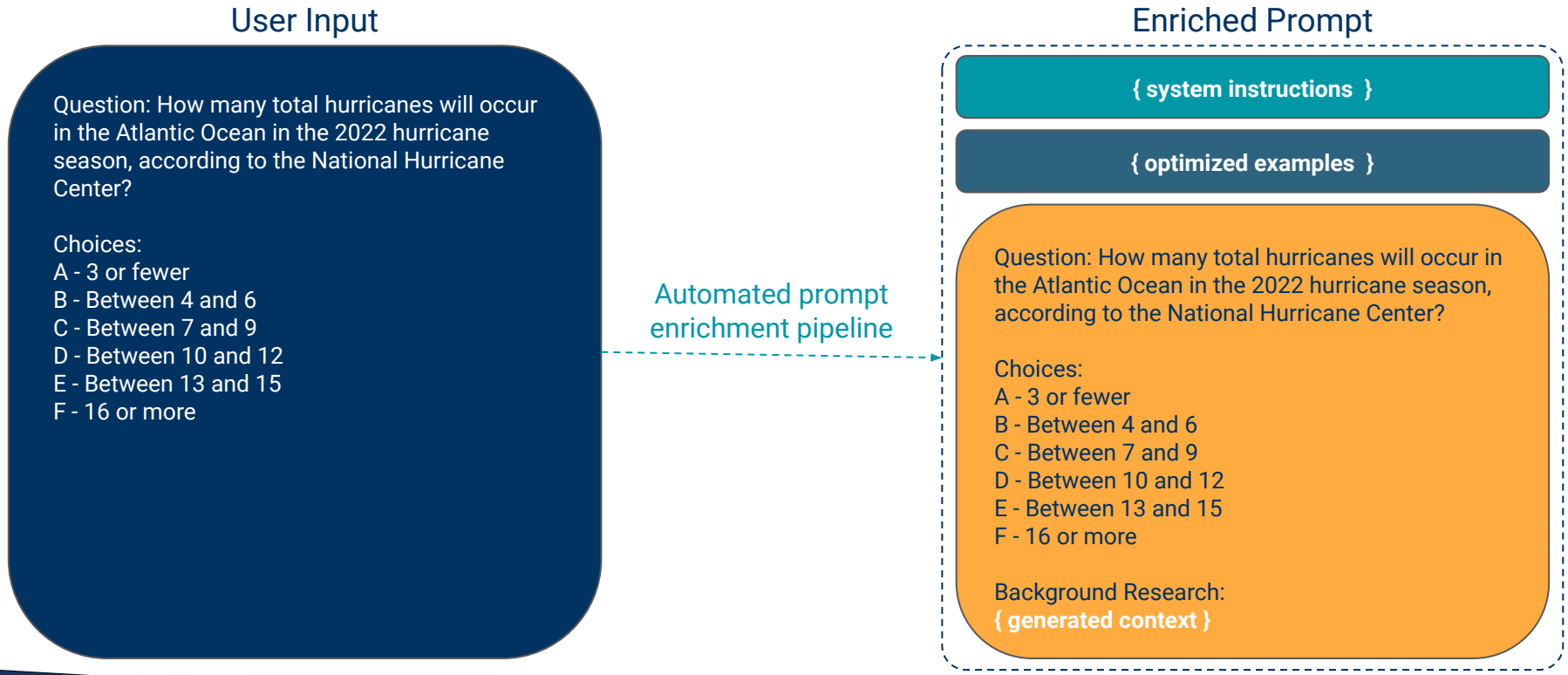
Predicting the Future with LLMs

Predicting future events is both hugely valuable and incredibly difficult. Our goal is to **use large language models to make forecasting easier**. We do this by leveraging existing knowledge embeddings and carefully designed prompts–all without the need for fine-tuning.

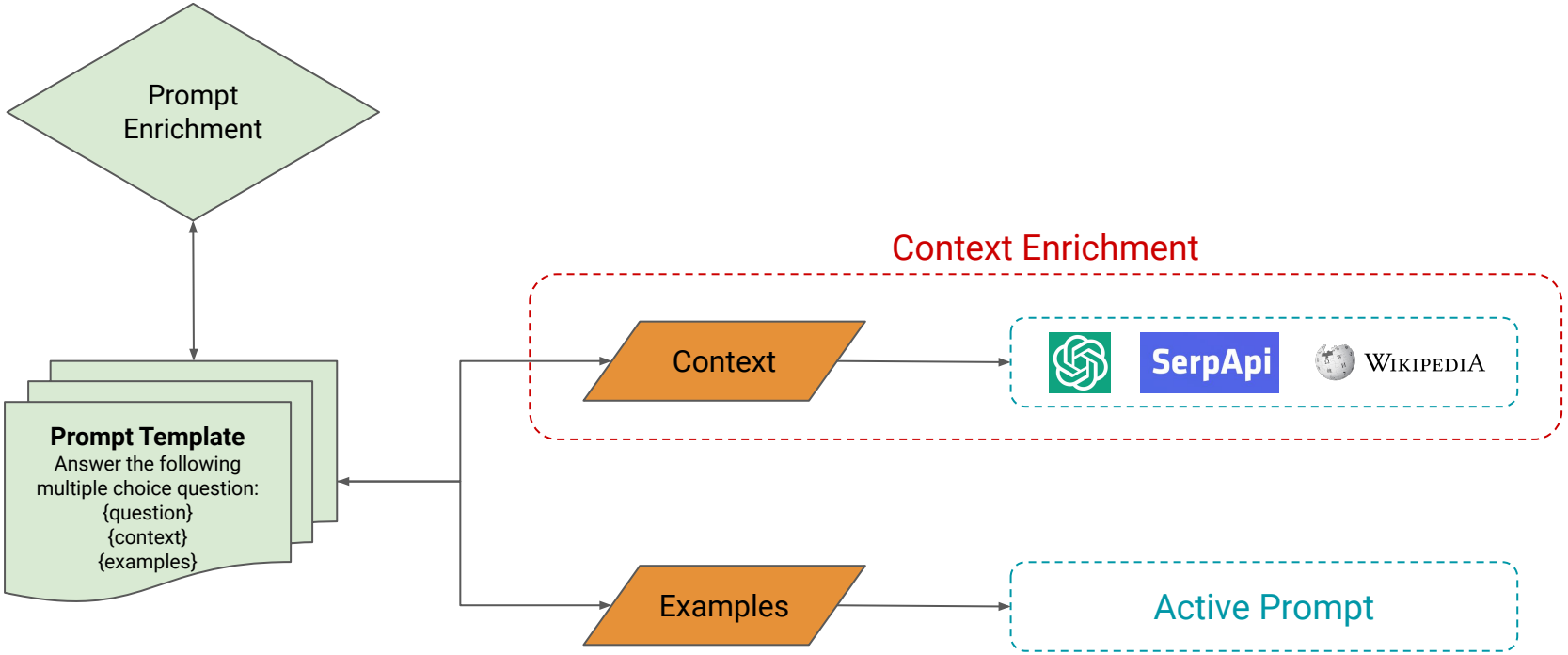
End-to-End LLM Forecasting Pipeline



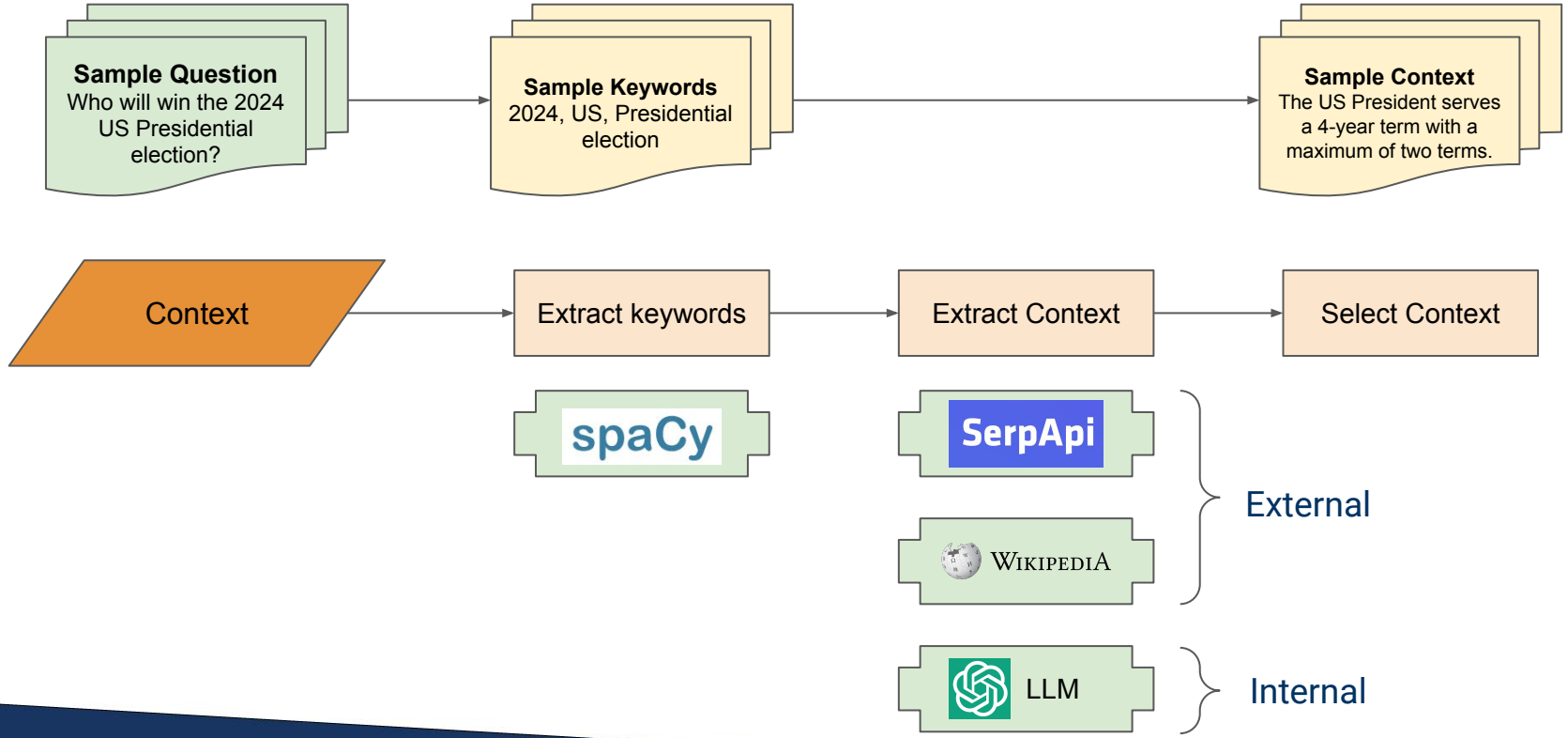
What is Prompt Enrichment?



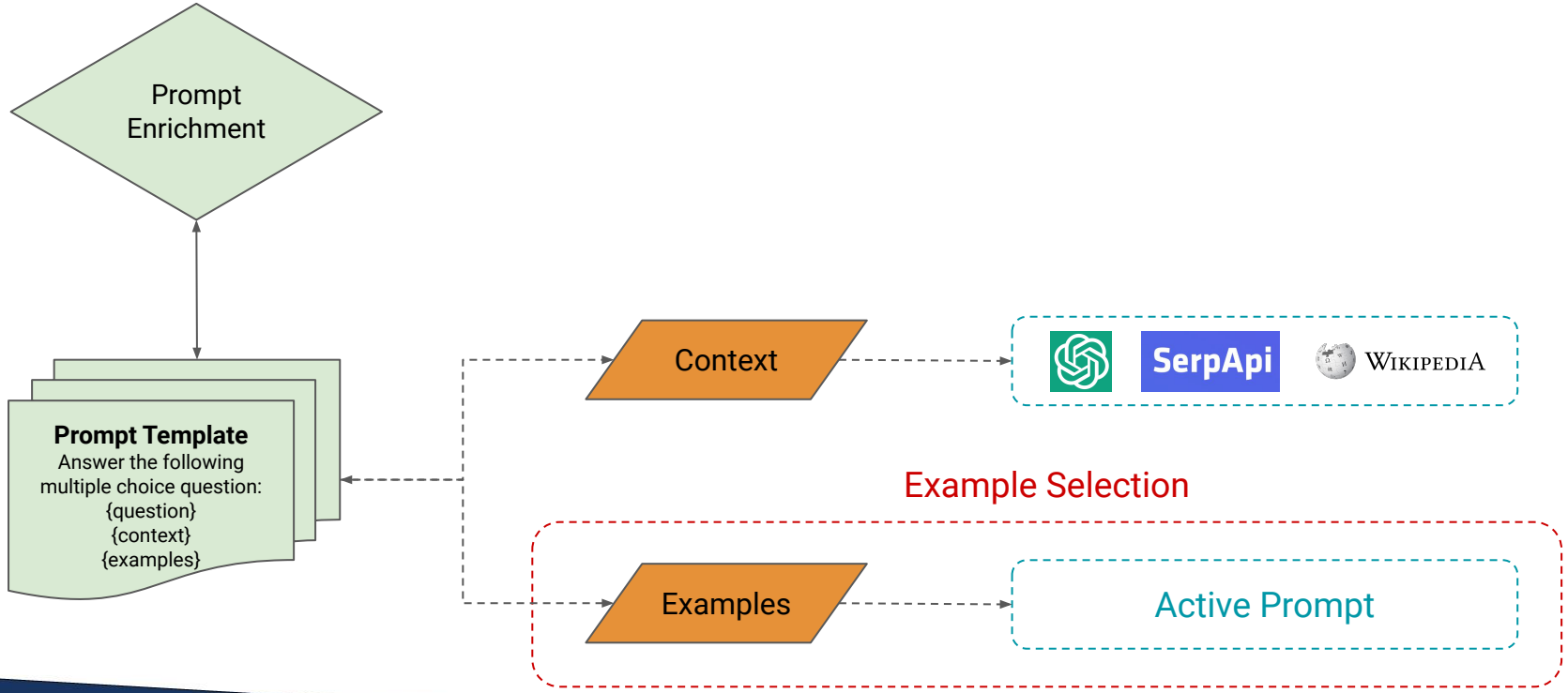
Prompt Enrichment - Context



Prompt Enrichment - Context

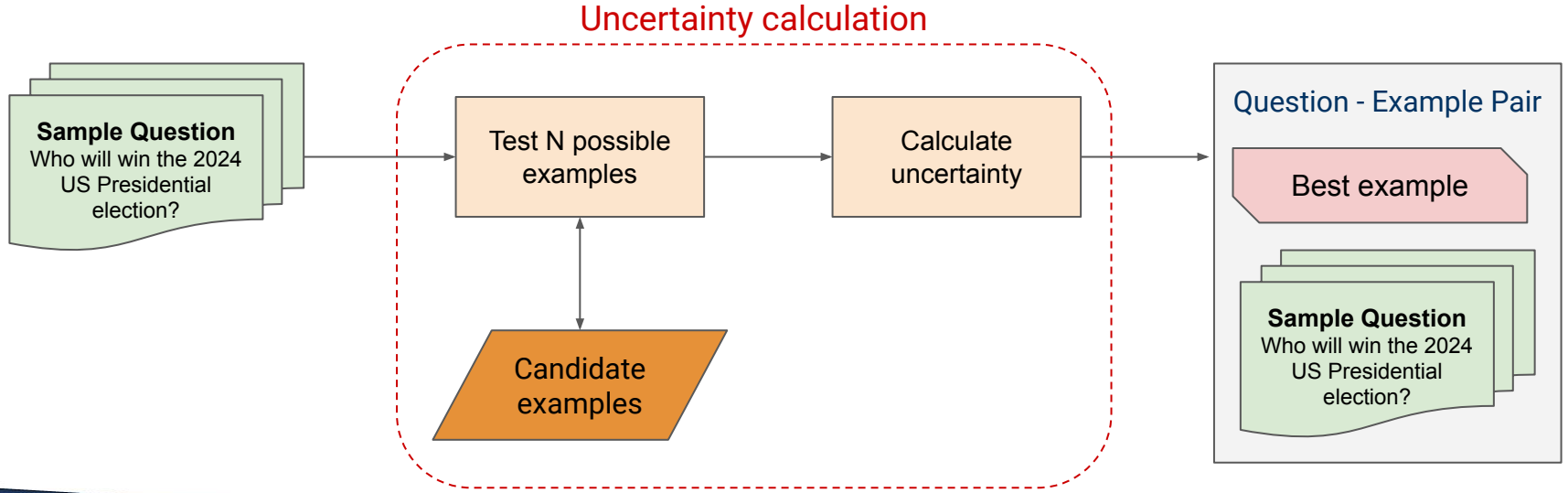


Prompt Enrichment - Example Selection

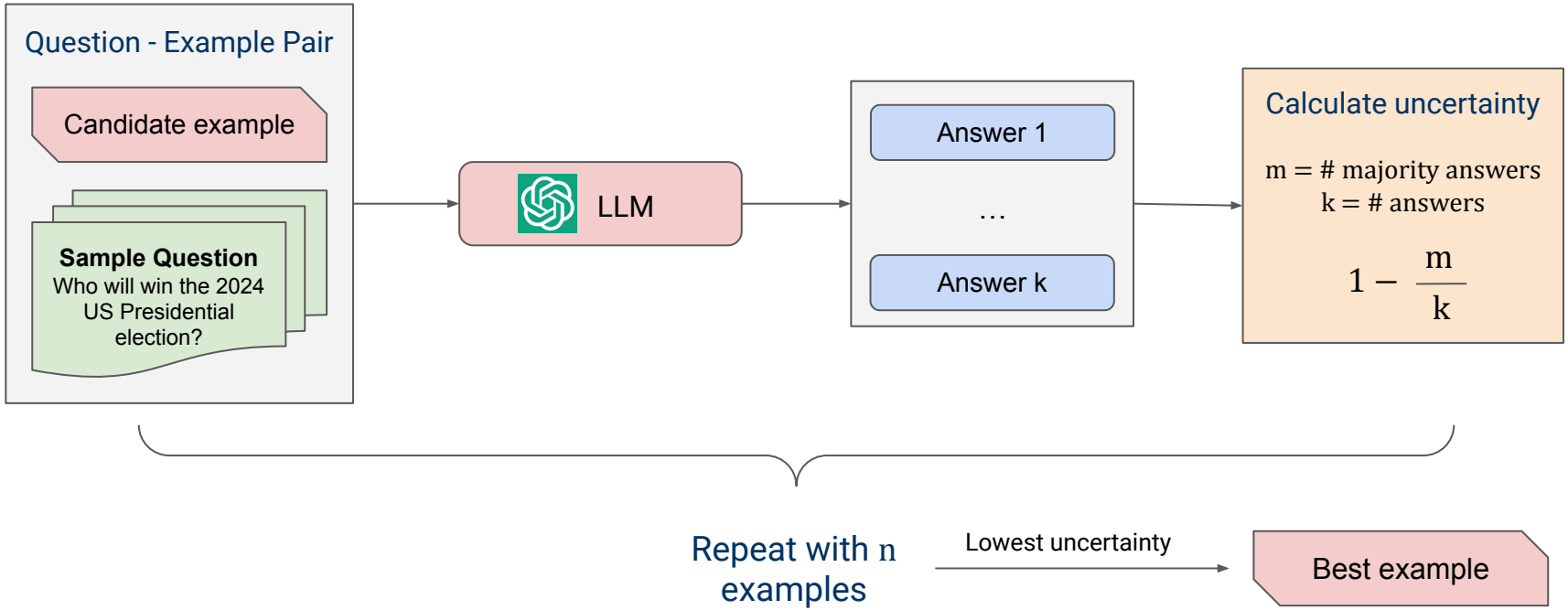


Prompt Enrichment - Active Prompt

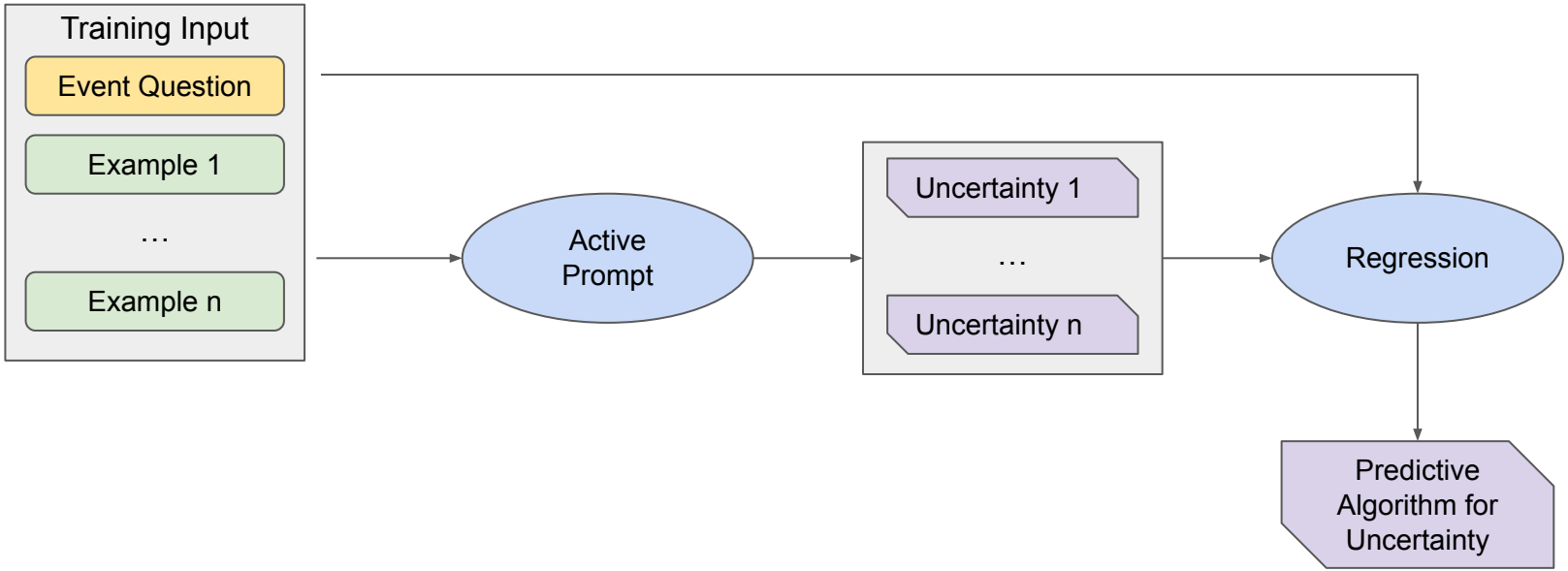
Active prompt lets us **select the best example** for a given question, based on how it **reduces model uncertainty**.



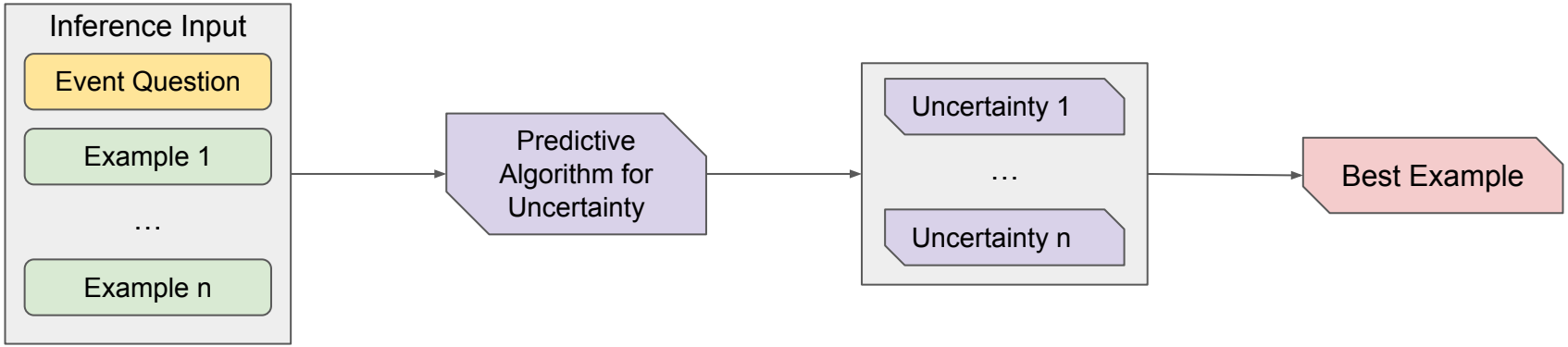
Prompt Enrichment - Active Prompt



Prompt Enrichment - Active Prompt



Prompt Enrichment - Active Prompt

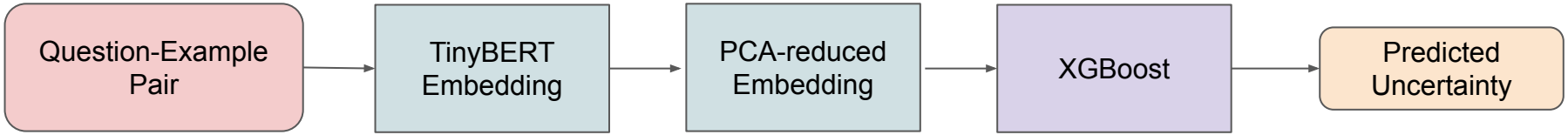


Prompt Enrichment - Active Prompt

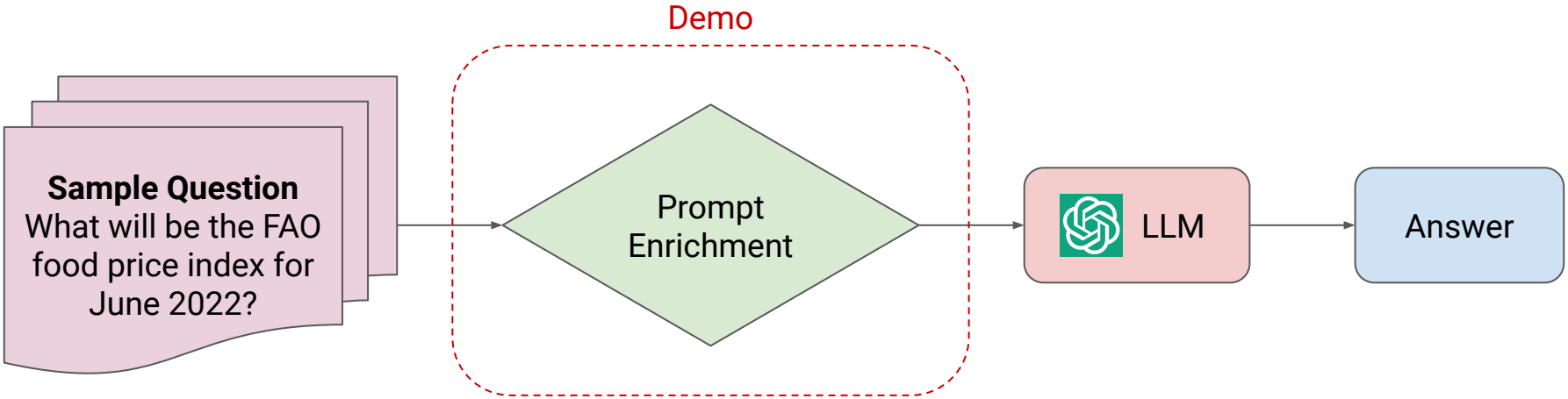
Central Tech Challenge for Regressor:

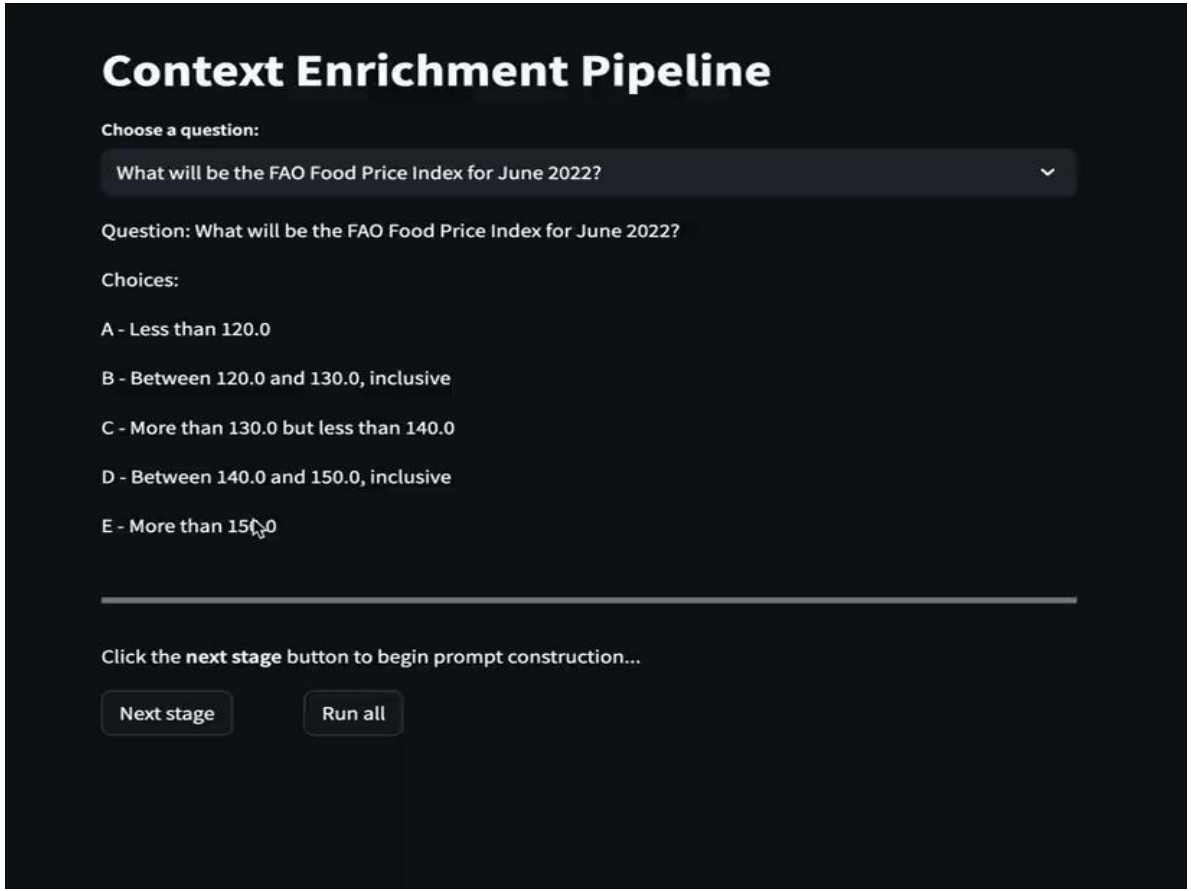
- High dimensionality of input data demands complex model
- Low data volume demands simpler model

Solution:



End-to-End LLM Forecasting Pipeline - Demo




Context Enrichment Pipeline
Choose a question:
What will be the FAO Food Price Index for June 2022?
Question: What will be the FAO Food Price Index for June 2022?
Choices:
A - Less than 120.0
B - Between 120.0 and 130.0, inclusive
C - More than 130.0 but less than 140.0
D - Between 140.0 and 150.0, inclusive
E - More than 150.0
Click the next stage button to begin prompt construction...
Next stage
Run all

Evaluation Dataset

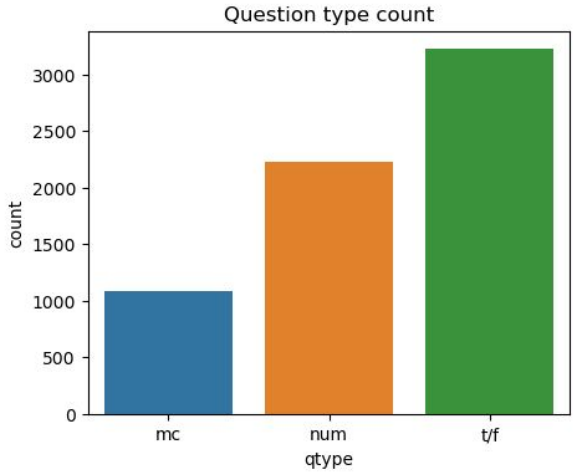
Sample Question:

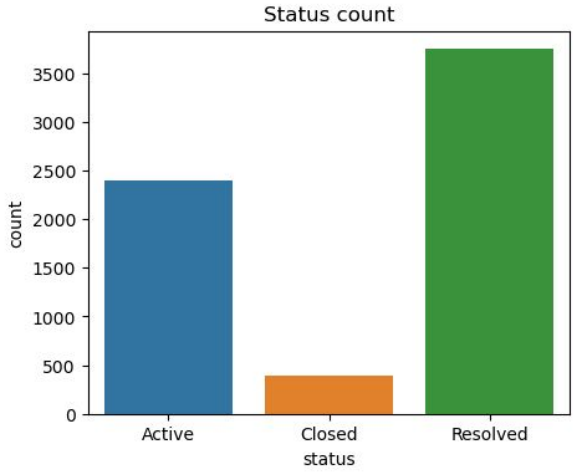
What will the end-of-day closing value for the dollar against the renminbi be on 1 January 2016?

Choices:

[‘Less than 6.30', 'Between 6.30 and 6.35, inclusive', 'More than 6.35 but less than 6.40', '6.40 or more']

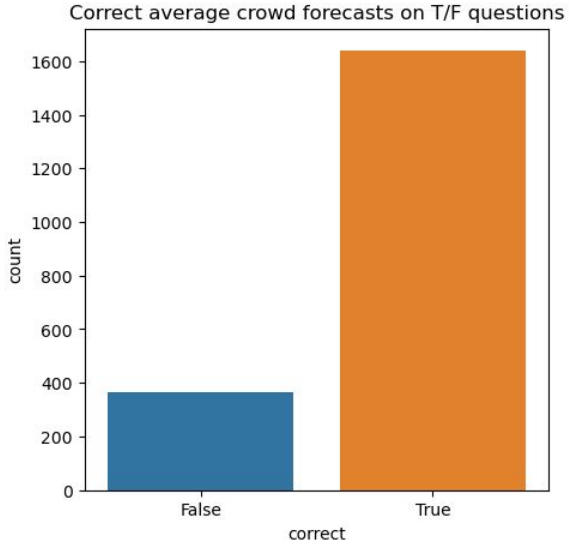
Answer: D

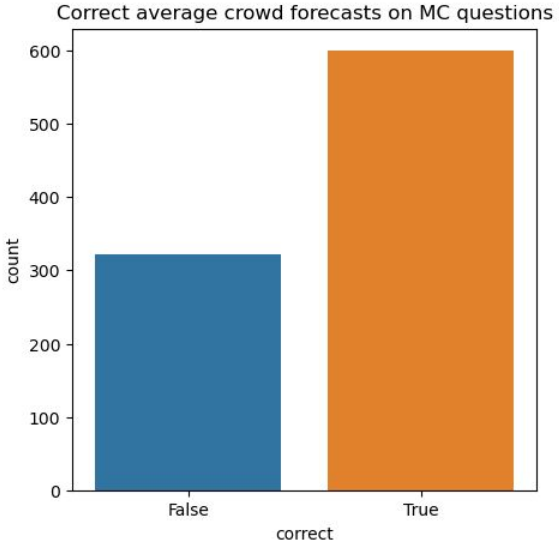




Evaluation Metric

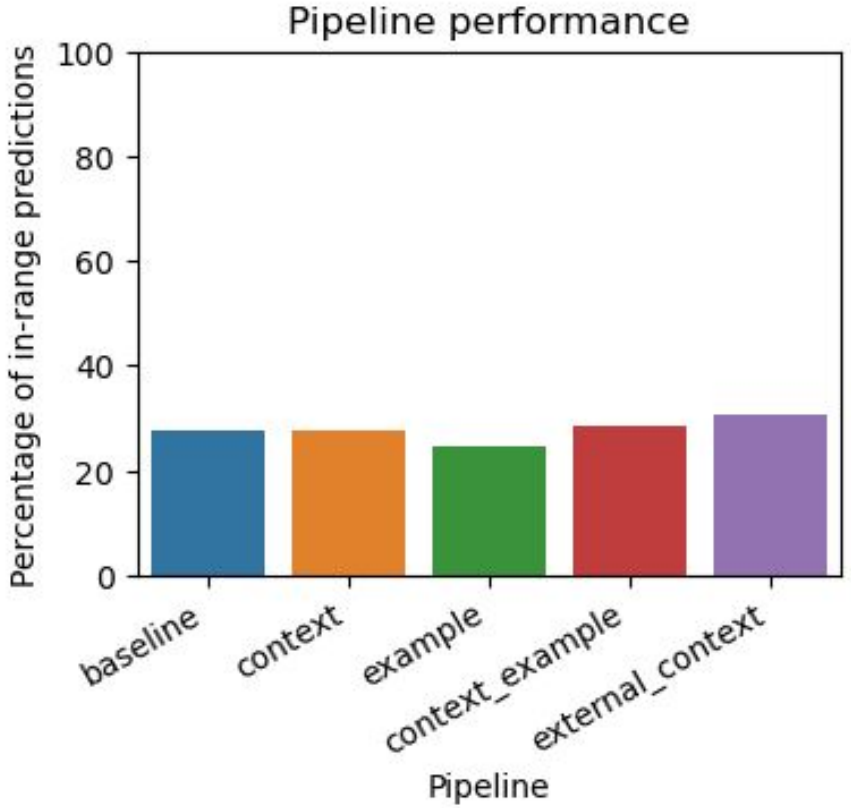
Crowd forecasts are an **imperfect truth proxy**, but still a useful target for our pipeline.





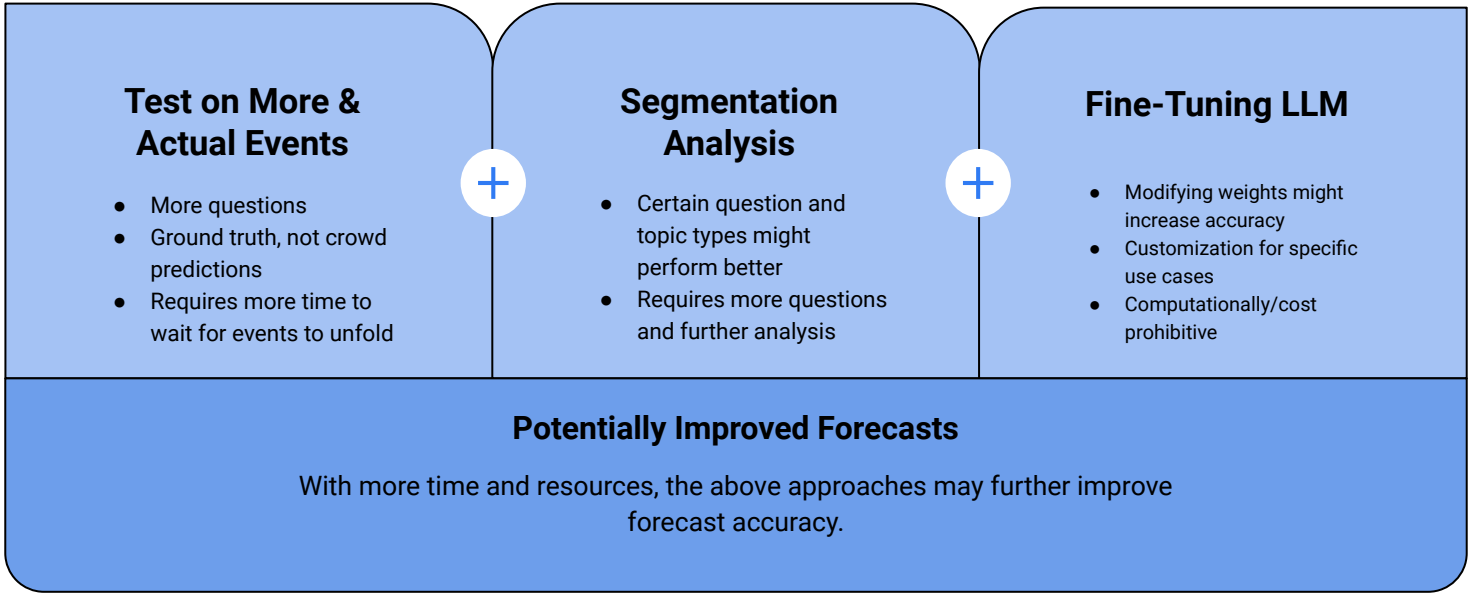
Pipeline Performance

Pipeline	% in-range predictions
Baseline	27.55%
Context	27.55%
External context	30.61%
Example	24.49%
Context + example	28.57%



Future Work

Future endeavors without current time or cost constraints could improve forecasting results.



Conclusion

- Minimal increases in forecasting performance when prompt engineering for **additional context.**
  - Larger sample needed for effective active learning model.
  - Target isn't reliable given the crowd's predictions are not always accurate.
- Putting forecasting aside, we learned the power of **LLMs**, **prompt engineering**, and **automated active learning** given minimal user queries.

Questions?