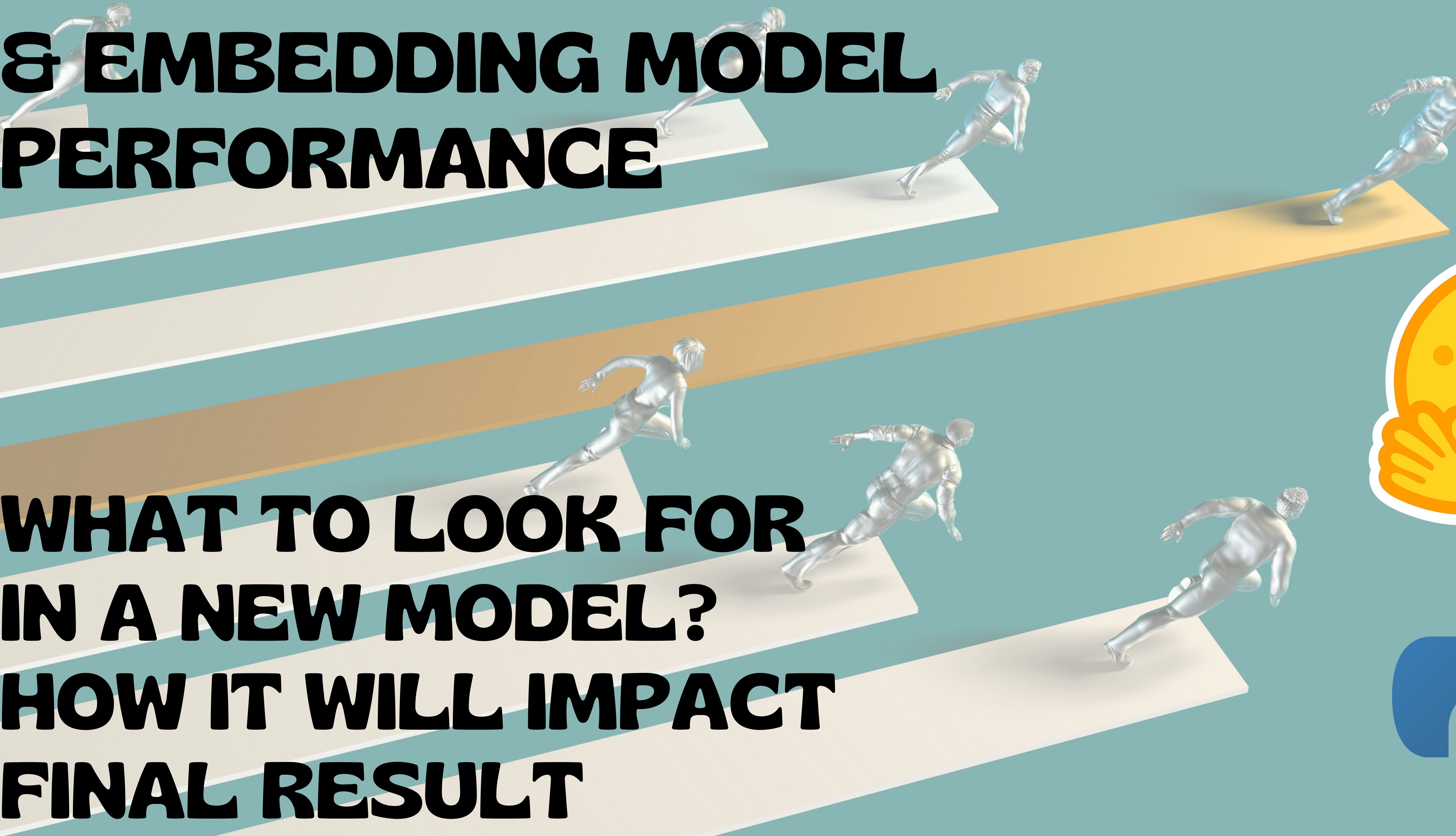
HTTPS://GITHUB.COM/INSIGHTBUILDER

# DEEP DIVE INTO LLM & EMBEDDING MODEL PERFORMANCE



## WHAT TO LOOK FOR IN A NEW MODEL? HOW IT WILL IMPACT FINAL RESULT

# CHALLENGE SOLVED : BY LEADERBOARD

- NEW TECHNOLOGY SOLVES OUR CHALLENGES, AND THIS TECHNOLOGY COSTS MONEY AND EFFORT. LEADERBOARDS HELPS US TO DECIDE WHERE TO SPEND OUR MONEY AND CAPITAL
- IT INTRODUCES US TO NEW CONCEPTS THAT WILL HELP US IN DESIGNING BETTER SYSTEMS. LEARNING THEM INSIDE OUT WILL LEAD TO BETTER RESULTS
- PROVIDES CONCRETE OBJECTIVES AND THE DIRECTION FOR THE FURTHER IMPROVEMENTS
- HELP IN DECISION MAKING AND PROVIDING THE DATA TO THE STAKE HOLDERS FOR INVESTMENTS
- CREATES HEALTHY COMPETITION AMONG THE DEVELOPERS, AND OPENS UP OPPORTUNITY FOR IMPROVEMENT

# LEADER BOARDS IN QUESTION

- MTEB: MASSIVE TEXT EMBEDDING BENCHMARK
- RAFT : REAL WORLD ANNOTATED FEW-SHOT TASKS
- OPEN LLM LEADERBOARD
- AUTOEVALUATE LEADERBOARDS

EACH TASKS HAVE THEIR OWN BENCHMARKS. EACH OF THESE MODELS CAN BE TESTED WITH AVAILABLE DATASETS IN THE HUB

AI ORGS CREATE THEIR OWN BENCHMARKS, FOR EXAMPLE ELEUTHERAI HAS 4 BENCHMARK AS HARNESSES

HUGGINGFACE HUB PROVIDES EASY INTERACE TO WORK WITH THE BENCHMARKS THROUGH ITS EVALUATE LIBRARY. THE TRANSFORMERS MODELS CAN BE EASILY LOADED INTO PIPELINE FOR TESTING

# OPEN LLM BENCHMARK : SIGNIFICANCE

THERE ARE 4 BENCHMARKS TO RANK THE TEXT-GENERATION / GENERATIVE AI LLMS

- AI2 REASONING CHALLENGE (ARC): SCIENCE QUESTIONS FROM GRADE SCHOOL
- HELLASWAG(10 SHOT): TEST OF COMMON SENSE WHICH SOTA MODELS STRUGGLE
- MMLU(5 SHOT): MULTI-TASK ACCURACY. TEST CONTAINS 57 TASKS RANGING FROM MATH TO LAW AND MORE
- TRUTHFUL QA (0 SHOT): TEST WHETHER MODEL IS TRUTHFUL IN ITS TEXT GENERATION

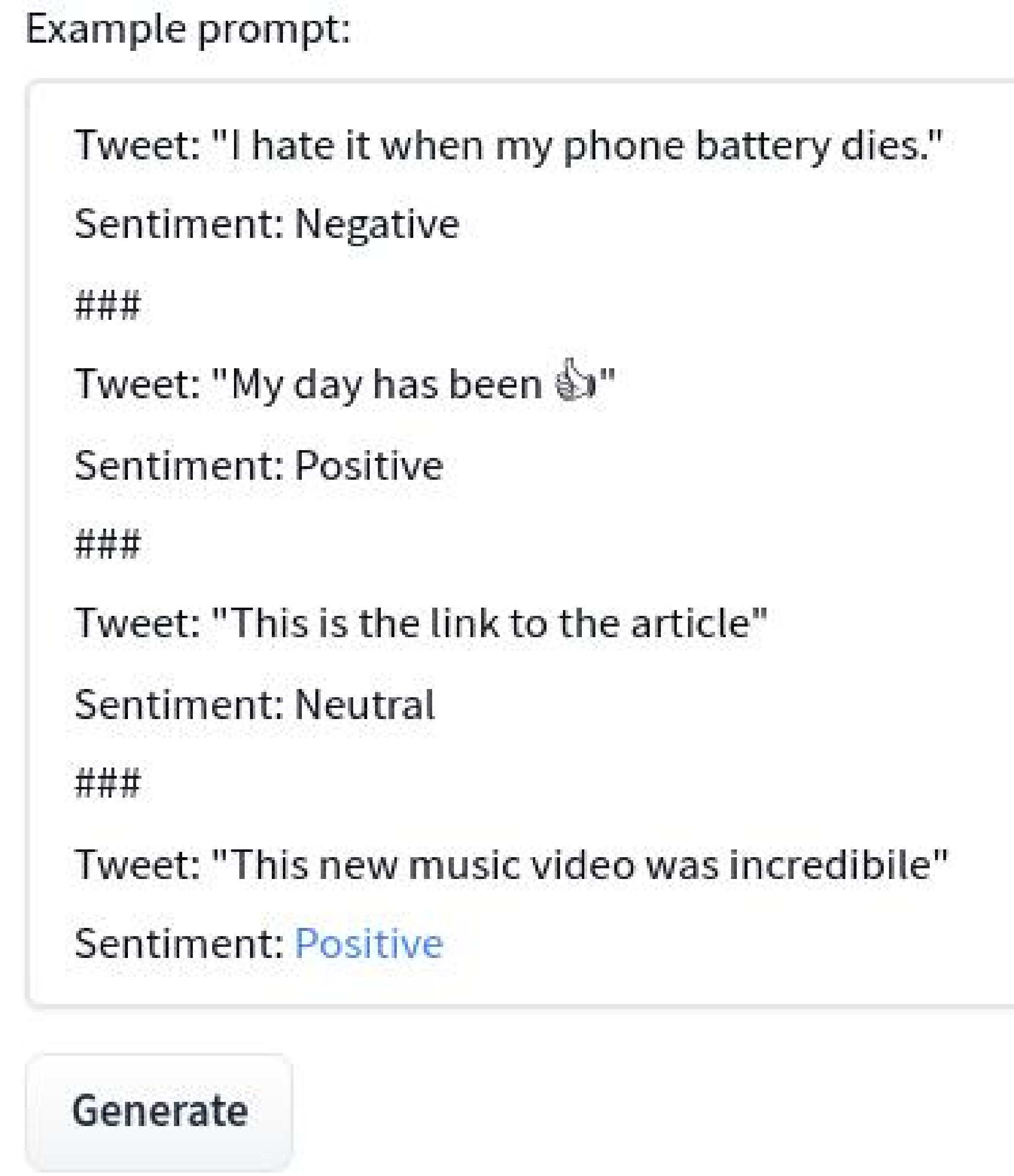
THE ABOVE IS ONLY THE TEST SELECTED BY OPEN LLM LEADER BOARD, ELEUTHER AI HAS LARGE NUMBER OF TASKS (200 +) CODED ALREADY

HTTPS://GITHUB.COM/ELEUTHERAI/LM-EVALUATION-HARNESS/BLOB/MASTER/DOCS/TASK_TABLE.MD

ELEUTHER AI FRAMEWORK CAN BE USED WITH PAID LLM LIKE OPENAI, TEXTSYNTH, AND GOOSE AI

# WHAT ARE 25, 5, & FEW SHOTS???

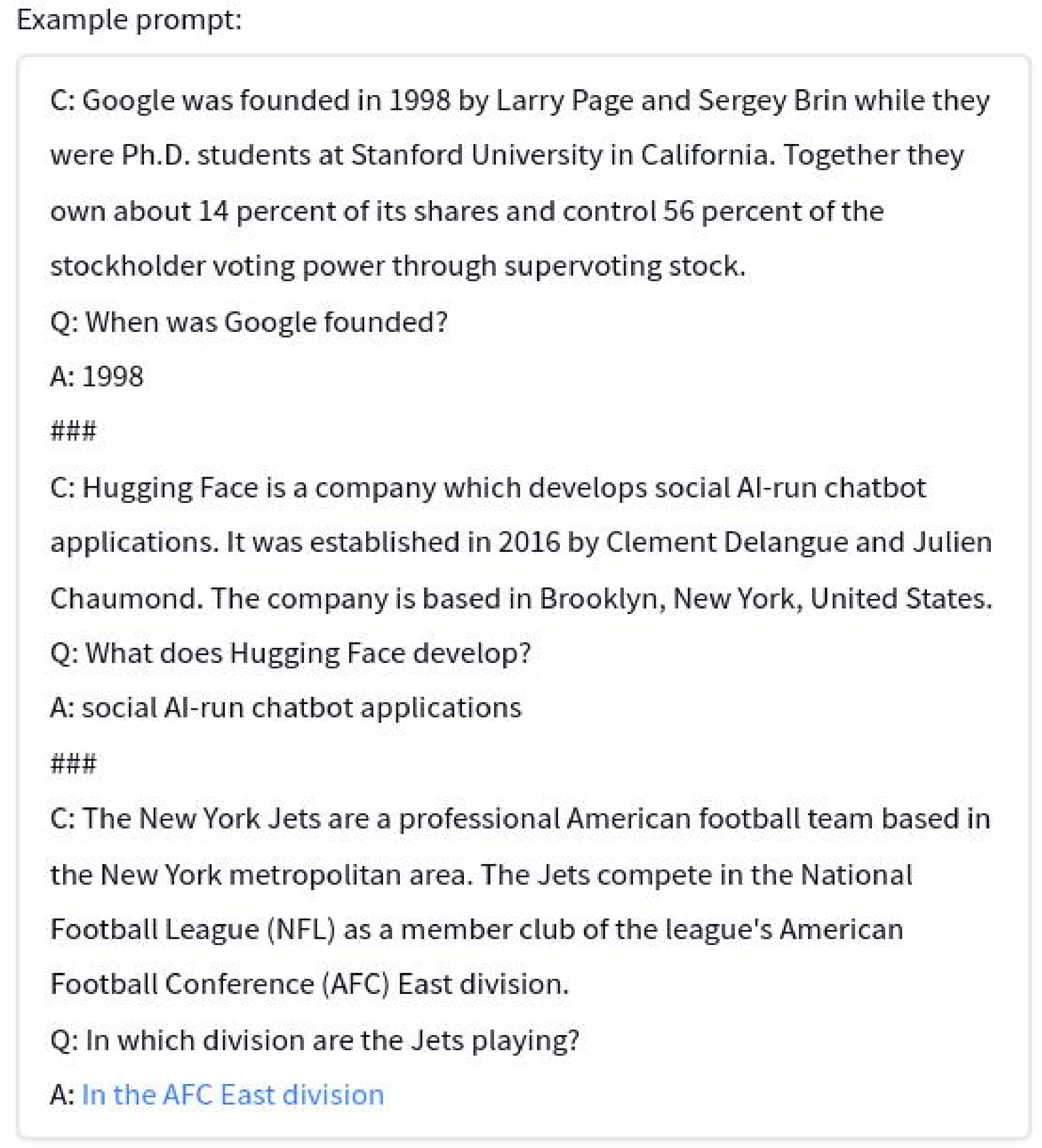
## FEW SHOT SENTIMENT ANALYSIS



Example prompt:

Tweet: "I hate it when my phone battery dies."
Sentiment: Negative
###
Tweet: "My day has been 👍"
Sentiment: Positive
###
Tweet: "This is the link to the article"
Sentiment: Neutral
###
Tweet: "This new music video was incredibile"
Sentiment: Positive

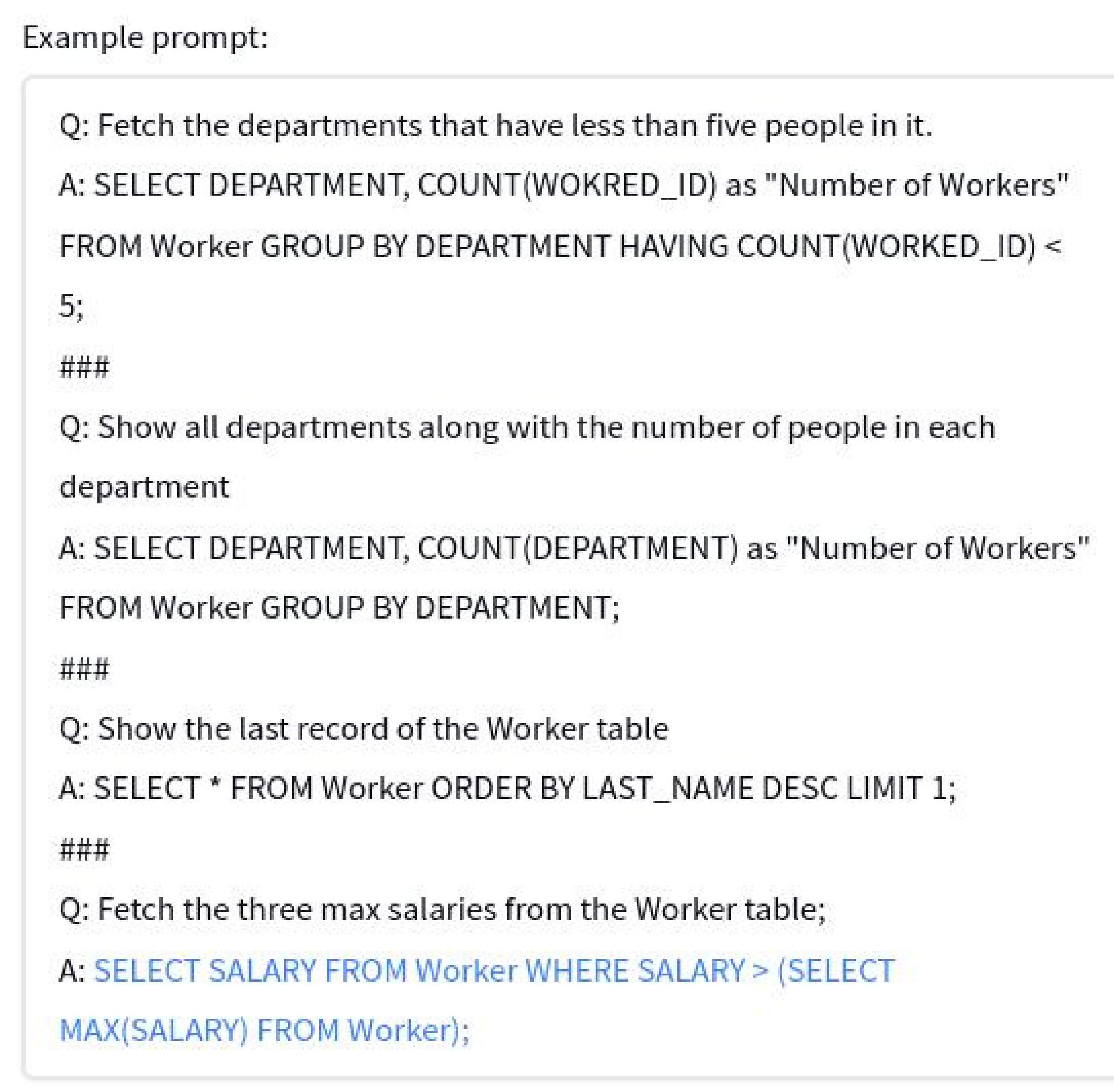
## FEW SHOT QUESTION ANSWERING



Example prompt:

C: Google was founded in 1998 by Larry Page and Sergey Brin while they were Ph.D. students at Stanford University in California. Together they own about 14 percent of its shares and control 56 percent of the stockholder voting power through supervoting stock.
Q: When was Google founded?
A: 1998
###
C: Hugging Face is a company which develops social AI-run chatbot applications. It was established in 2016 by Clement Delangue and Julien Chaumond. The company is based in Brooklyn, New York, United States.
Q: What does Hugging Face develop?
A: social AI-run chatbot applications
###
C: The New York Jets are a professional American football team based in the New York metropolitan area. The Jets compete in the National Football League (NFL) as a member club of the league's American Football Conference (AFC) East division.
Q: In which division are the Jets playing?
A: In the AFC East division

## FEW SHOT SQL GENERATION



Example prompt:

Q: Fetch the departments that have less than five people in it.
A: SELECT DEPARTMENT, COUNT(WOKRED_ID) as "Number of Workers" FROM Worker GROUP BY DEPARTMENT HAVING COUNT(WORKED_ID) < 5;
###
Q: Show all departments along with the number of people in each department
A: SELECT DEPARTMENT, COUNT(DEPARTMENT) as "Number of Workers" FROM Worker GROUP BY DEPARTMENT;
###
Q: Show the last record of the Worker table
A: SELECT * FROM Worker ORDER BY LAST_NAME DESC LIMIT 1;
###
Q: Fetch the three max salaries from the Worker table;
A: SELECT SALARY FROM Worker WHERE SALARY > (SELECT MAX(SALARY) FROM Worker);

**LLM IS PROVIDED WITH TASK DESCRIPTION, EXAMPLE PROMPTS AND FOLLOWUP ANSWERS. AT THE END THE QUESTION/ PROMPT THAT WE WANT THE LLM TO REPLY IS INCLUDED.**

**WHEN YOU PROVIDE MORE EXAMPLES , LLM UNDERSTANDS THE TASK AND TAKES THE END_SEQUENCE INTO ACCOUNT, WHICH ALLOWS US TO CONTROL THE GENERATED TEXT PRETTY WELL.**

**HTTPS://GITHUB.COM/INSIGHTBUILDER**

# HOW TO THINK ABOUT THE BENCH MARKS

- WHAT TASK THE TEST IS TRYING TO EVALUATE. DOES THAT TASK SCORE MATTERS FOR MY APPLICATION?
- AFTER A MODEL IS FINE-TUNED FOR A SPECIFIC TASK, THE IMPROVEMENT IN THE MODEL CAN BE OBSERVED
- TUNING WITH DIFFERENT QUALITY OF DATASETS WILL HAVE DIFFERENT RESULTS
- IN CASE OF EMBEDDING MODELS THE VECTORS RETURNED HAVE TO SUCCEED IN TASKS LIKE BI-TEXT MINING,CLASSIFICATION, CLUSTERING, RETRIEVAL, RERANKING, SUMMARIZING AND STS(SEMANTIC TEXT SIMILARITY)

# THANKS FOR WATCHING

♥ LIKE

SHARE

SUBSCRIBE

HTTPS://GITHUB.COM/INSIGHTBUILDER