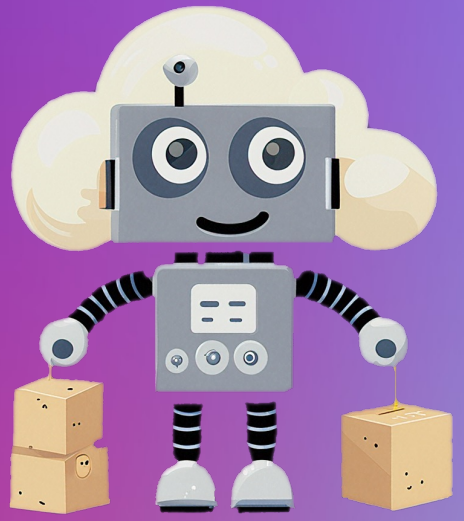
SES104

Evaluating Large Language Models

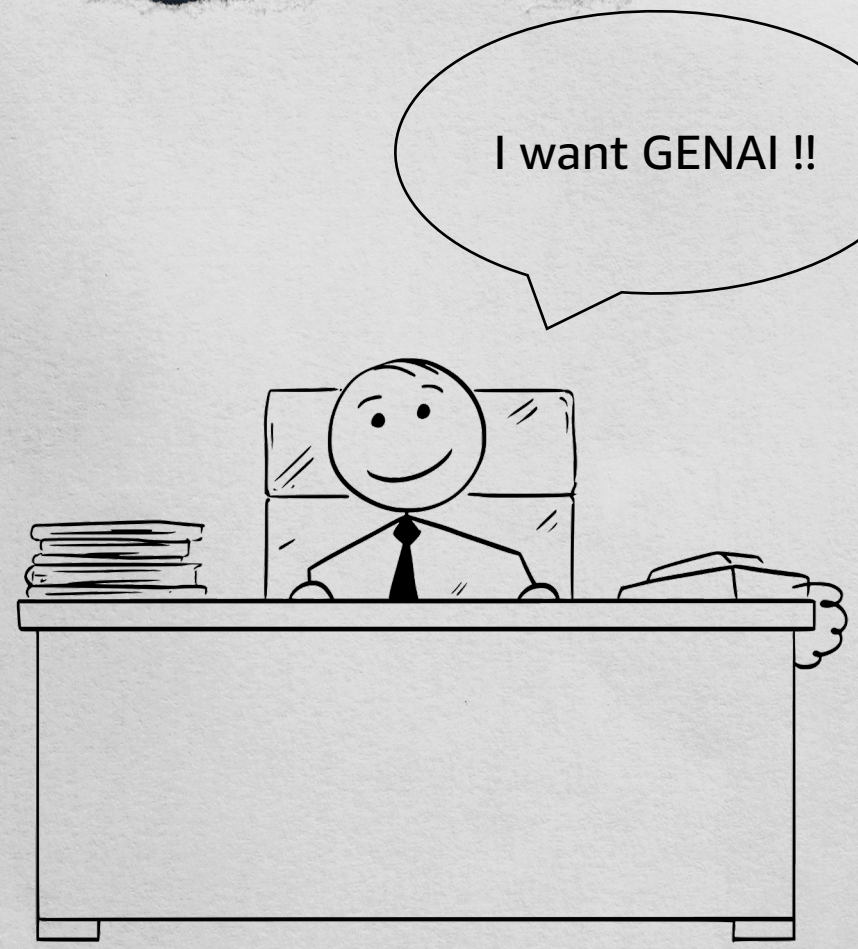
Practical Considerations and Open Challenges

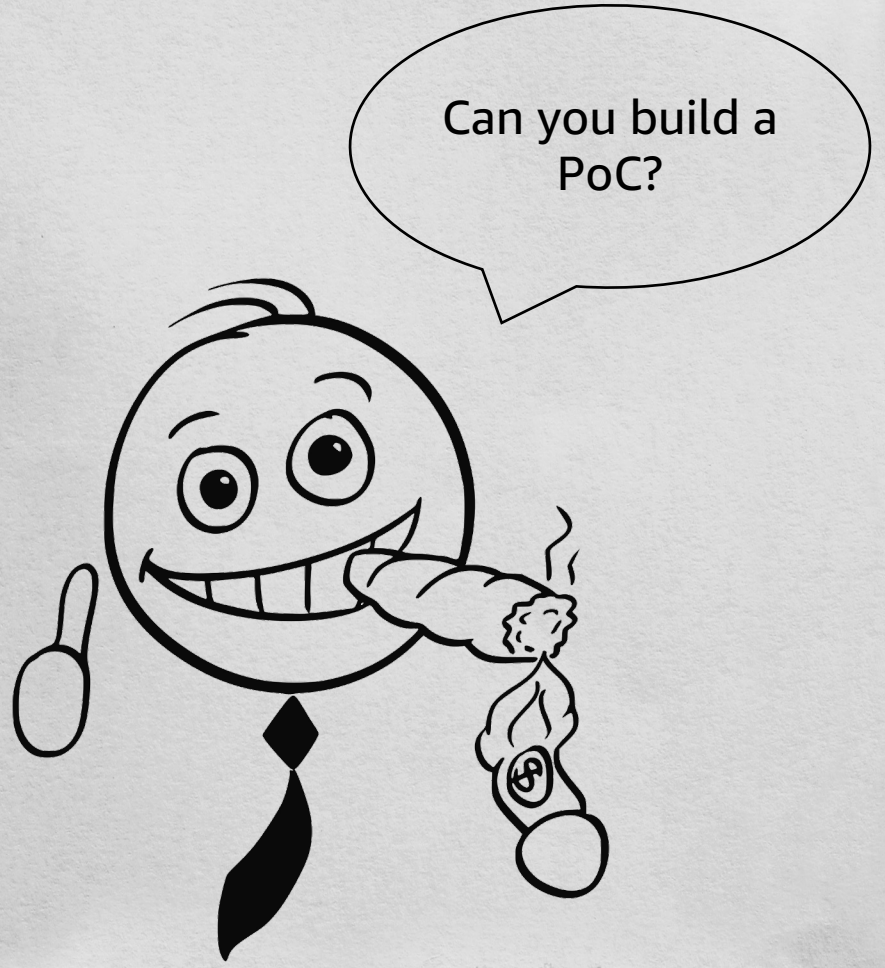
Lukas Wenzel
Solutions Architect, AWS

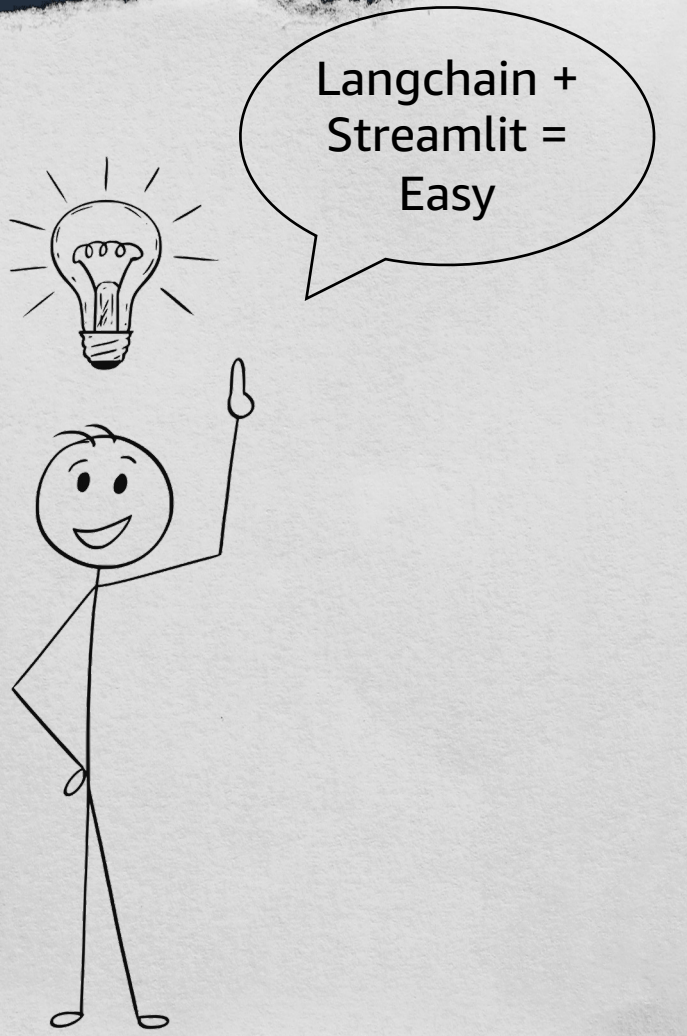
Johannes Langer
Solutions Architect – AI, AWS

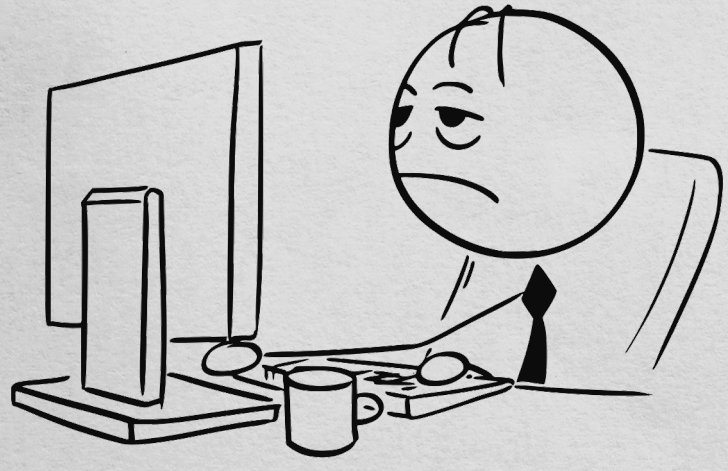


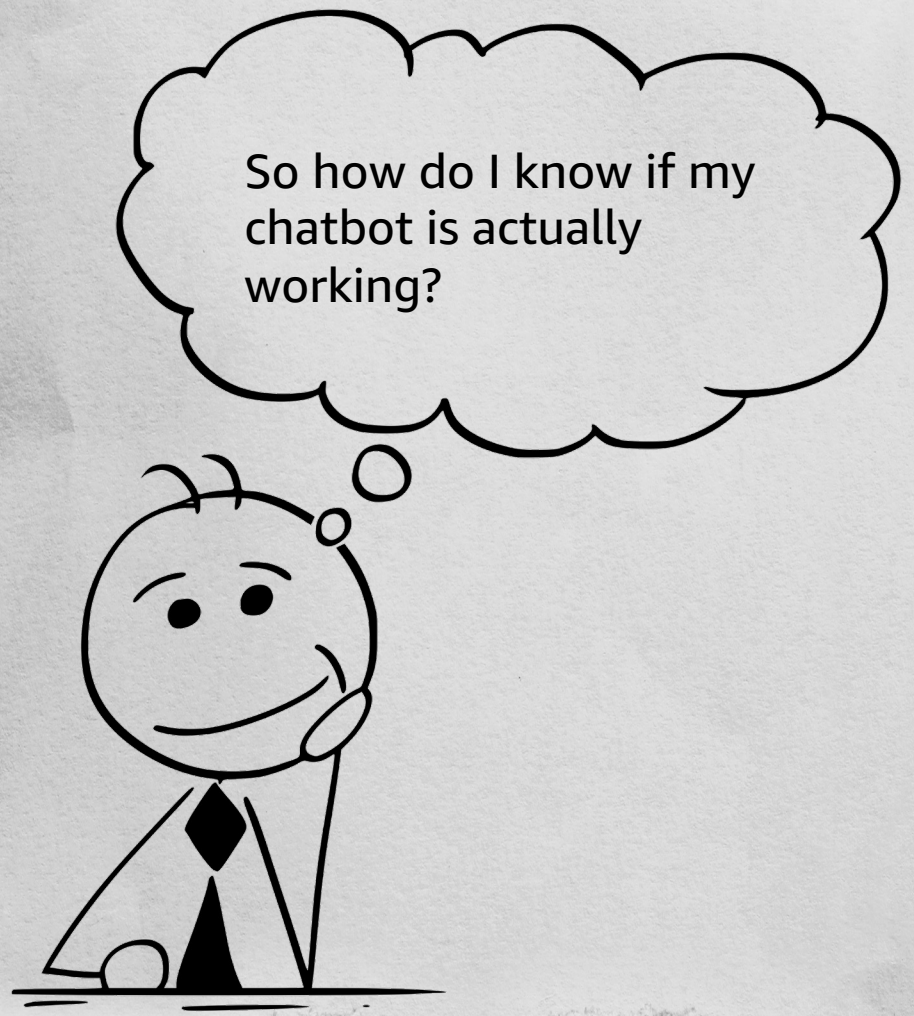
Bob the GenAI Builder



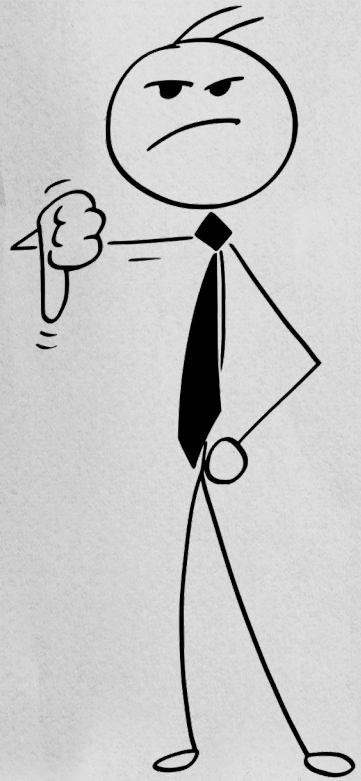






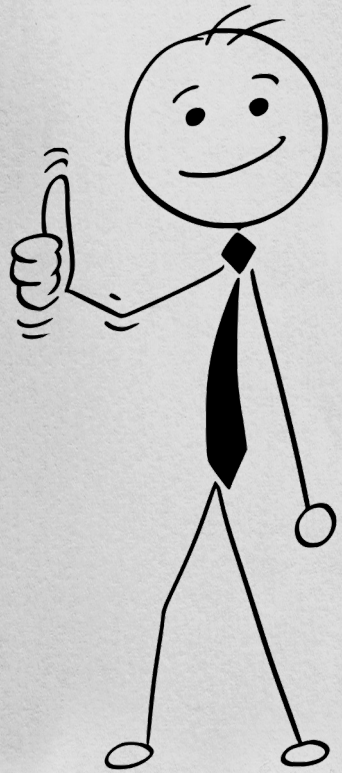


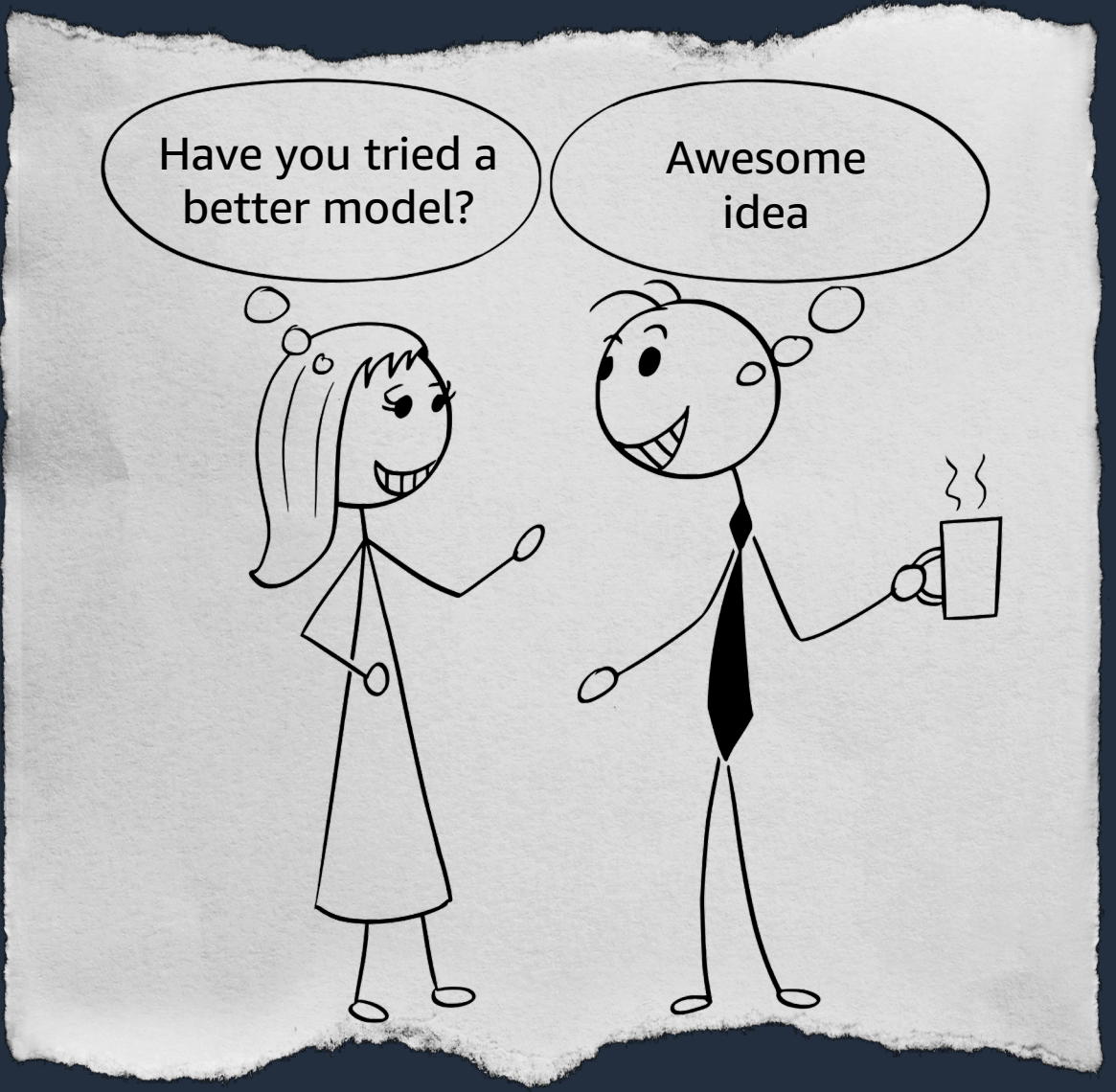
Quantative Evaluation?



Vibe Checks?

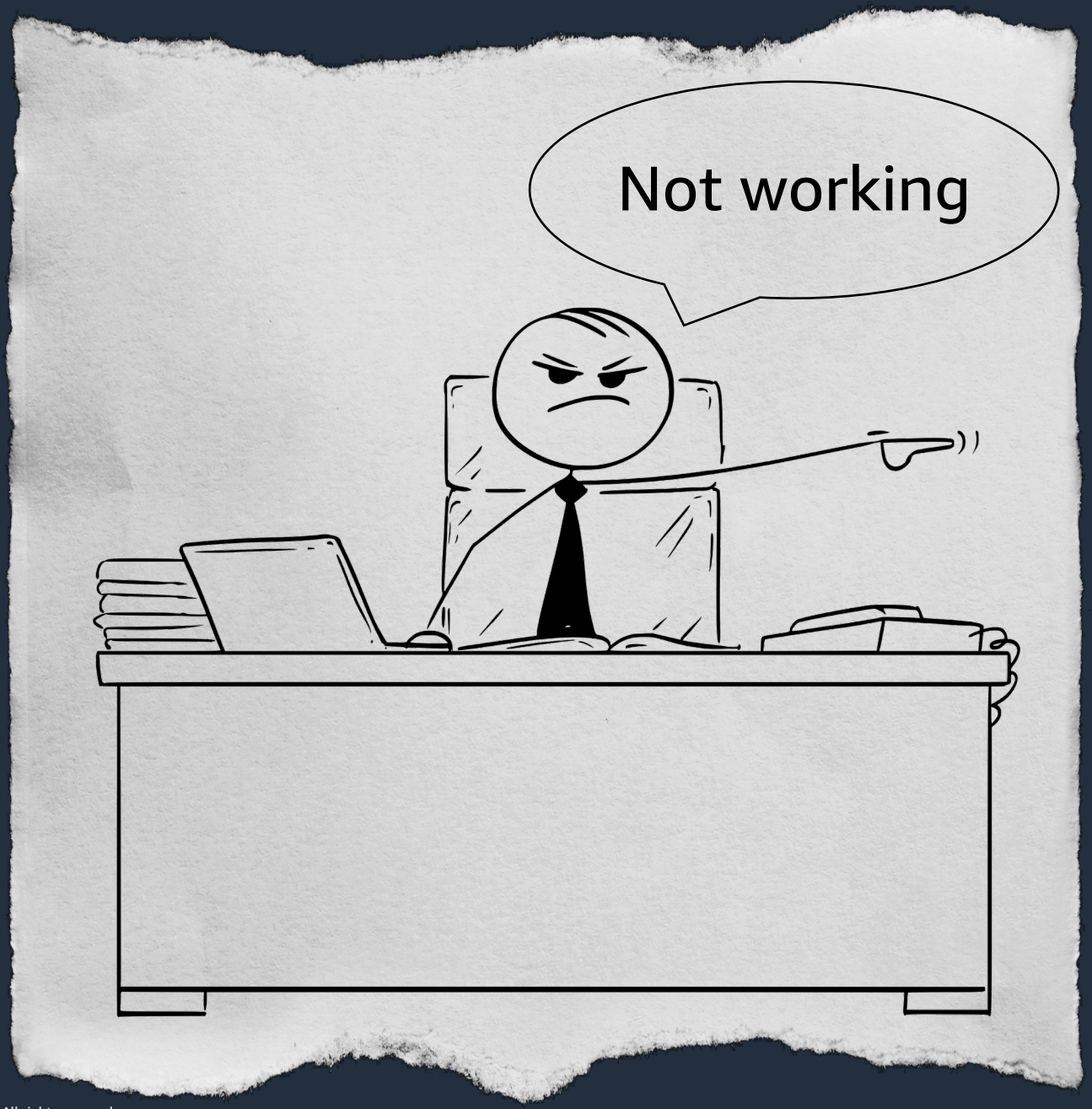
(a.k.a. I just try a few prompts and see if its working)

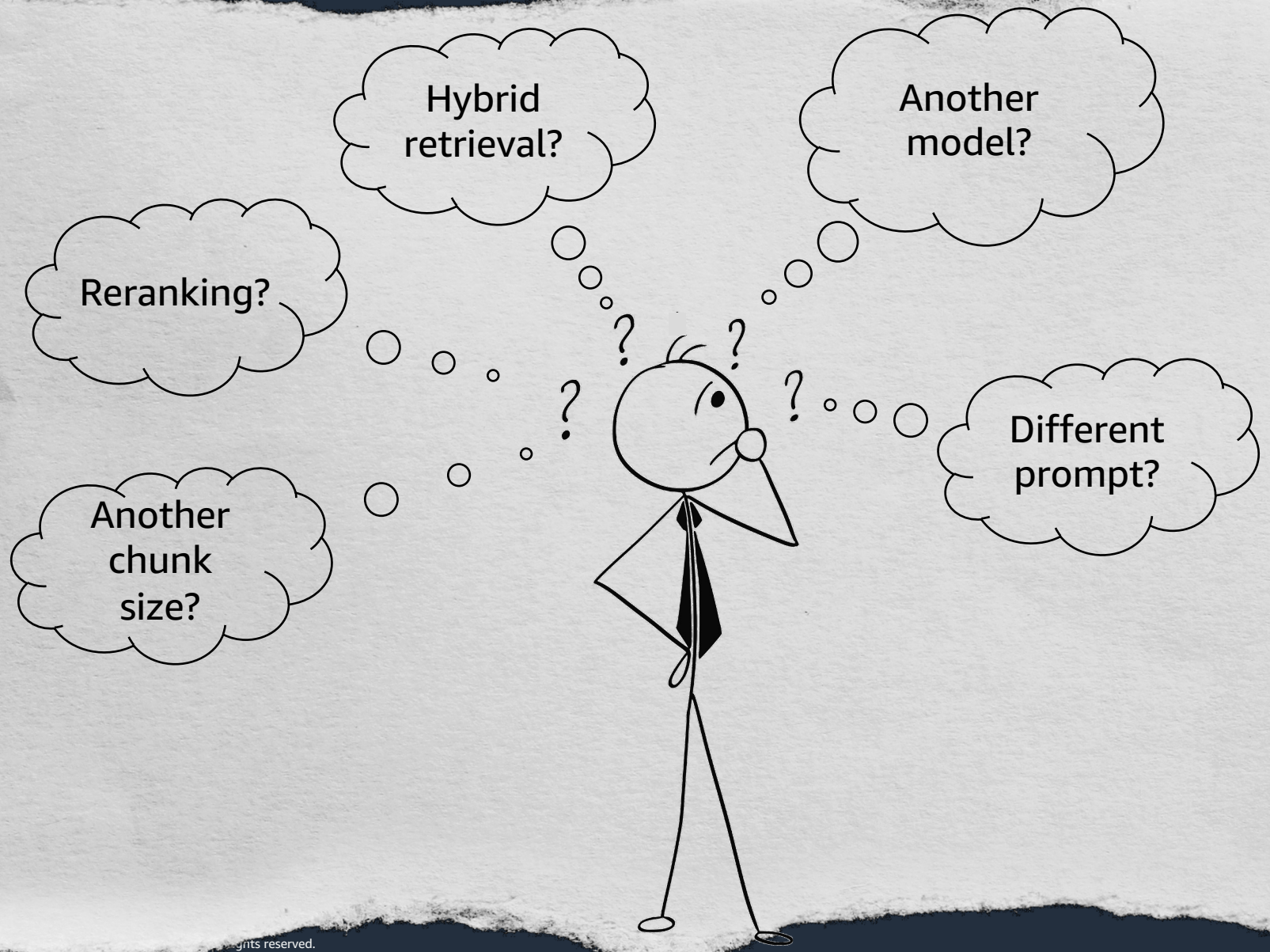




T	Model
🔶	abacusai/Smaug-72B-v0.1
🔶	ibivibiv/alpaca-dragon-72b-v1
💬	moreh/MoMo-72B-lora-1.8.7-DPO
🔶	cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_DPO_f16
🔶	cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_full_linear_DPO
💬	yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B

https://crfm.stanford.edu/helm/lite/v1.0.0/#/scenarios / https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard






NEED
HELP!!!

Why not pick the best model from leaderboard?

HELM Lite — Leaderboard | Models | Scenarios | Predictions | GitHub | Release: v1.0.0

Accuracy | Efficiency | General information

Model	Mean win rate	NarrativeQA - F1	NaturalQuestions (open-book) - F1	NaturalQuestions (closed-book) - F1	OpenbookQA - EM	MMLU - EM	MATH - Equivalent (CoT)
GPT-4 (0613)	0.962	0.768	0.79	0.457	0.96	0.735	0.802
GPT-4 Turbo (1106 preview)	0.834	0.727	0.763	0.435	0.95	0.699	0.857
Palmyra X V3 (72B)	0.821	0.706	0.685	0.407			
Palmyra X V2 (33B)	0.783	0.752	0.752	0.428			
PaLM-2 (Unicorn)	0.776	0.583	0.674	0.435			

HELM - Holistic Evaluation of Language Models

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande
💬	moreh/MoMo-72B-lora-1.8.7-DPO	78.55	70.82	85.96	77.13	74.71	84.06
💬	yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B	77.44	74.91	89.3	64.67	78.02	88.24
💬	moreh/MoMo-72B-lora-1.8.6-DPO	77.29	70.14	86.03	77.4	69	84.37
🔶	abacusai/Smaug-34B-v0.1	77.29	74.23	86.76	76.66	70.22	83.66
🔶	cloudyu/Truthful_DPO_TomGrc_FusionNet_34Bx2_MoE	77.28	72.87	86.52	76.96	73.28	83.19
🔶	yunconglong/13B_MATH_DPO	77.08	74.66	89.51	64.53	78.63	88.08
🔶	TomGrc/FusionNet_34Bx2_MoE	77.07	72.95	86.22	77.05	71.31	83.98
🔶	yunconglong/MoE_13B_DPO	77.05	74.32	89.39	64.48	78.47	88
💬	zhengr/MixTAO-7Bx2-MoE-Instruct-v7.0	76.55	74.23	89.37	64.54	74.26	87.77
🔶	cloudyu/Truthful_DPO_cloudyu_Mixtral_34Bx2_MoE_60B	76.48	71.25	85.24	77.28	66.74	84.29
💬	moreh/MoMo-72B-lora-1.8.4-DPO	76.23	69.62	85.35	77.33	64.64	84.14
🔶	TomGrc/FusionNet_7Bx2_MoE_14B	75.91	73.55	88.84	64.68	69.6	88.16

Huggingface - Open LLM Leaderboard

https://crfm.stanford.edu/helm/lite/v1.0.0/#/scenarios / https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

No evals

With evals

Eugene Yan
@eugeneyan

Teams with the most success using LLMs in prod prioritized evals & finetuning data

Automated evals enable experiments at scale and safe deployments by measuring improvement/regressions

Finetuning flywheels identify & correct defects and then finetune LLMs to get desired output

6:18 PM · Oct 29, 2023 · **13K** Views

6 11 81 30

Which metric should I use?

How to get test data?

How do I start ???

Is there a tool for this?

How to integrate this into my dev flow

Challenges with custom evaluations

Metrics

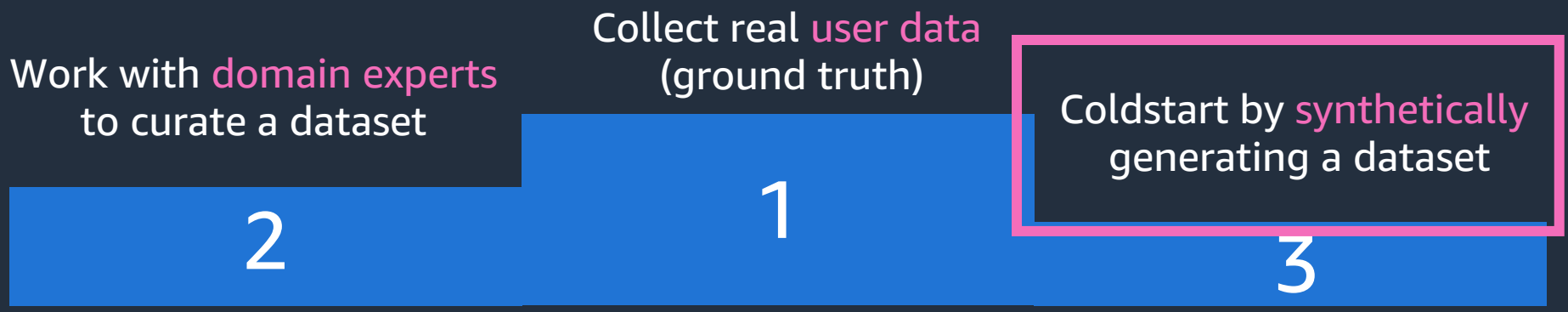
Tooling

Guidance

How to start

Acquiring a dataset for evaluation

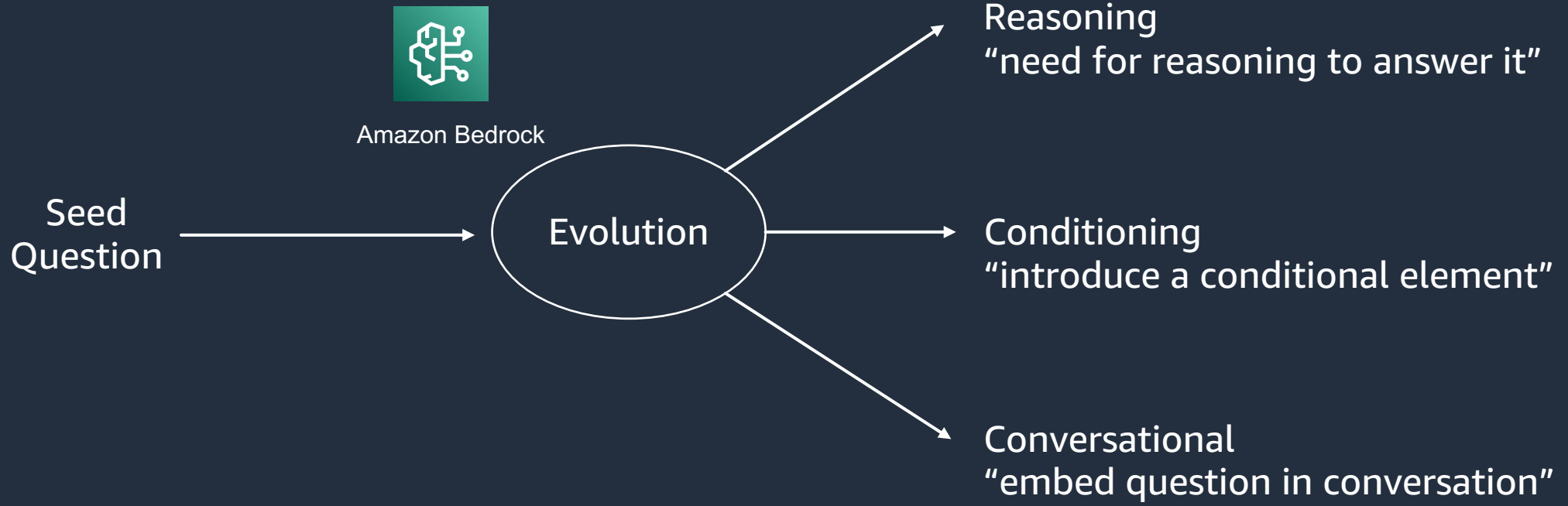
```
"inputRecord": {
    "prompt": "What does vitamin C serum do for skin?",
    "referenceResponse": "Vitamin C serum offers a range of benefits for the skin...
}
```



Create a synthetic dataset

generate_dataset_prompt= """
<role>You are an experienced QA Engineer [...] </role>
<task> Create a question that users could have about the following text [...]
</task>
<rules> Rule 1: The question should make sense to humans even without the given text [...] </rules>
<text> {text} <text>
<output_format>
{
question: “insert question here”
}
<output_format>
"""

Improve the dataset



Extend the dataset

	question	context	answer	question_type	episode_done
0	What technique improves the performance of lar...	- "We explore how generating a chain of though...	The technique that improves the performance of...	simple	True
1	What phenomenon is discussed in the paper rega...	- This paper instead discusses an unpredictabl...	The phenomenon discussed in the paper is the e...	reasoning	True
2	What is the purpose of chain-of-thought (CoT) ...	- Providing these steps for prompting demonstr...	The purpose of chain-of-thought (CoT) promptin...	simple	True
3	What is the performance of the largest fine-tu...	On the MathQA-Python dataset, the largest fine...	The performance of the largest fine-tuned mode...	simple	True
4	What is the accuracy increase of Zero-shot-CoT...	Experimental results demonstrate that our Zero...	The accuracy increase of Zero-shot-CoT on Mult...	reasoning	True

Source: https://docs.ragas.io/en/latest/concepts/testset_generation.html

Key takeaways

- Use synthetic data creation gets you **started quickly**
- **Oversee** creation, especially in **early iterations** of datasets
- Creation quality is **very sensitive to prompting,** implement **use case specific best practices** and **start small**
- Continuously **improve the dataset**, in best case with **user data**

LLM Evaluation methods

Operational and business considerations

- Quality
- Cost
- Latency
- Security and Data Privacy
- Licensing
- Capacity to operate the model

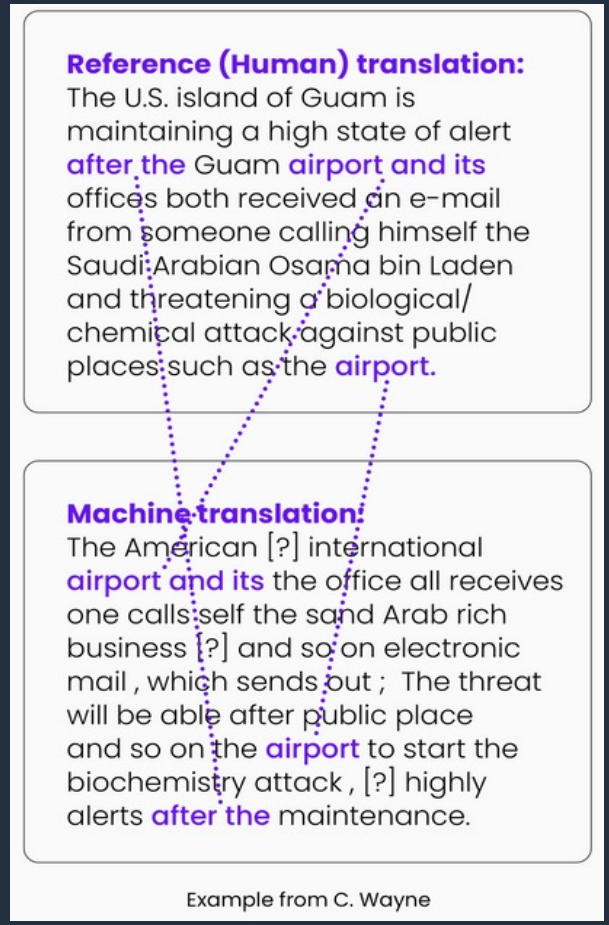
Anthropic models	Price per 1,000 input tokens
Claude Instant	$0.00080
Claude	$0.00800

Amazon Bedrock

Amazon SageMaker

What about traditional NLP metrics?

```
•[6]: from langchain.evaluation import ExactMatchStringEvaluator

      evaluator = ExactMatchStringEvaluator()

      evaluator.evaluate_strings(
          prediction="Generative AI Deep Dive Days Munich 24",
          reference="generative AI Dive Deep Days Munich 24",
      )

 [6]: {'score': 0}
```



Source: https://blog.modernmt.com/understanding-mt-quality-bleu-scores/

Abstractive nature of LLM tasks and generations

Who is Barack Obama?

LLM 1

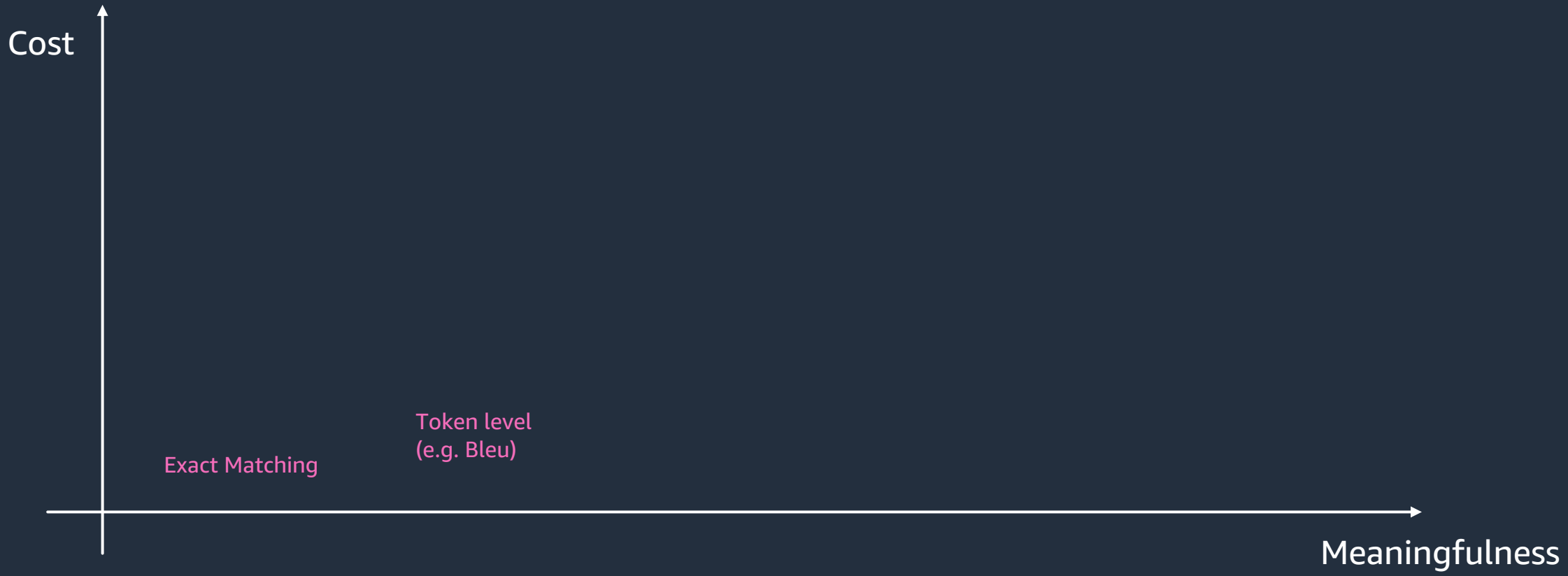
"Barack Obama is the 44th President of the United States, serving from 2009 to 2017."

Vs.

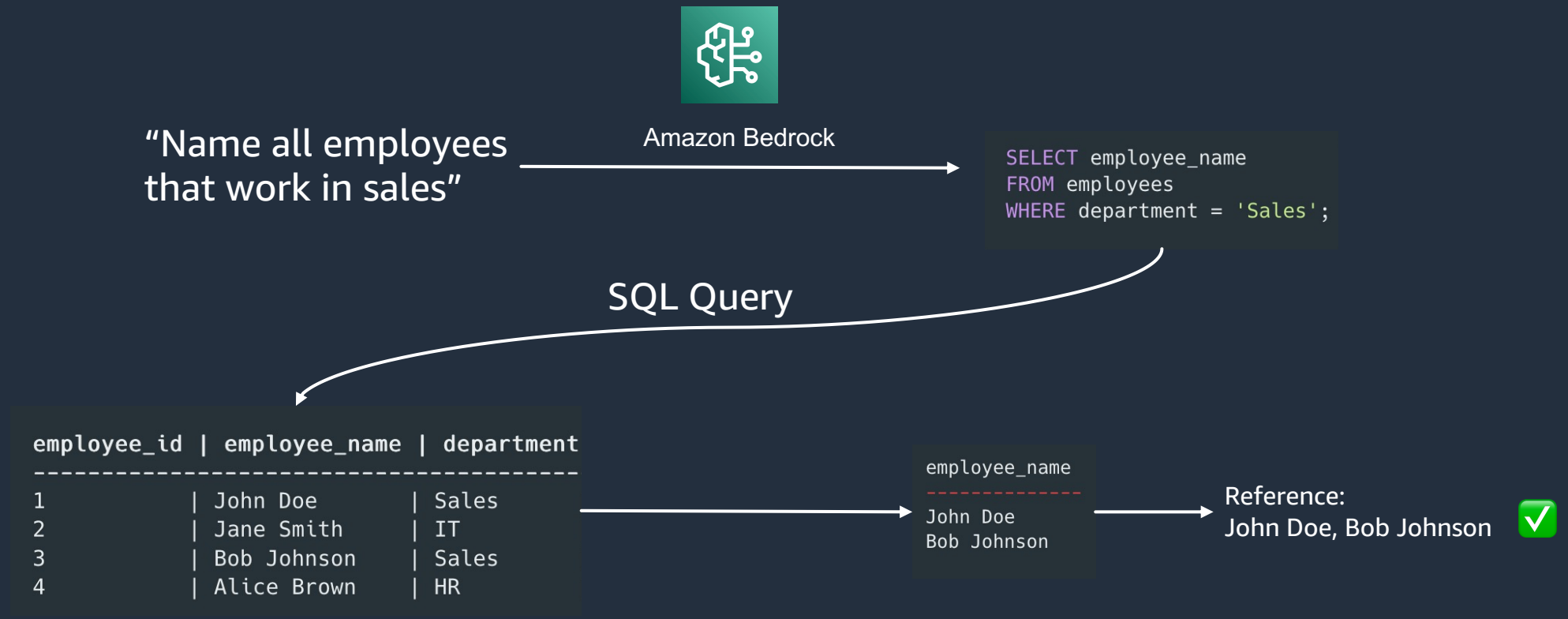
LLM 2

"Barack Obama, a Harvard Law graduate and former Illinois Senator, made history as the first African American President of the United States […]."

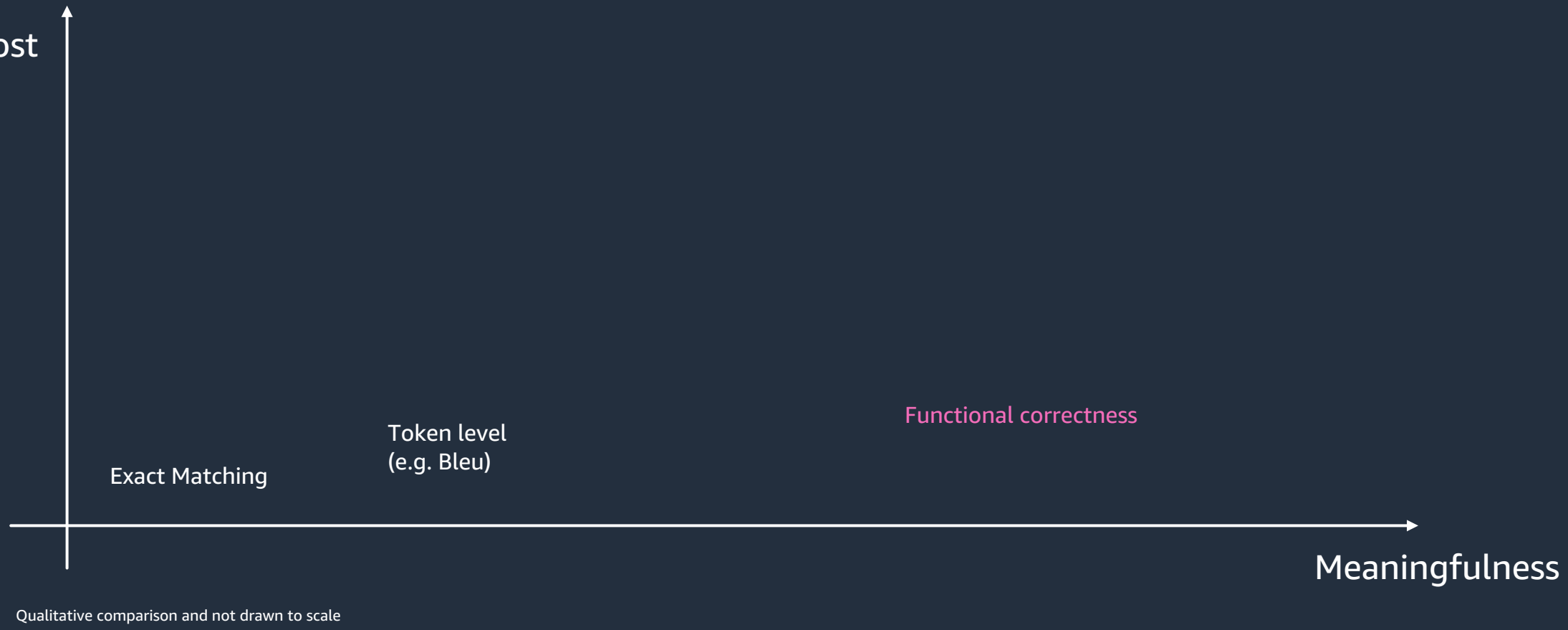
Evaluation methods for LLM systems:



Functional correctness

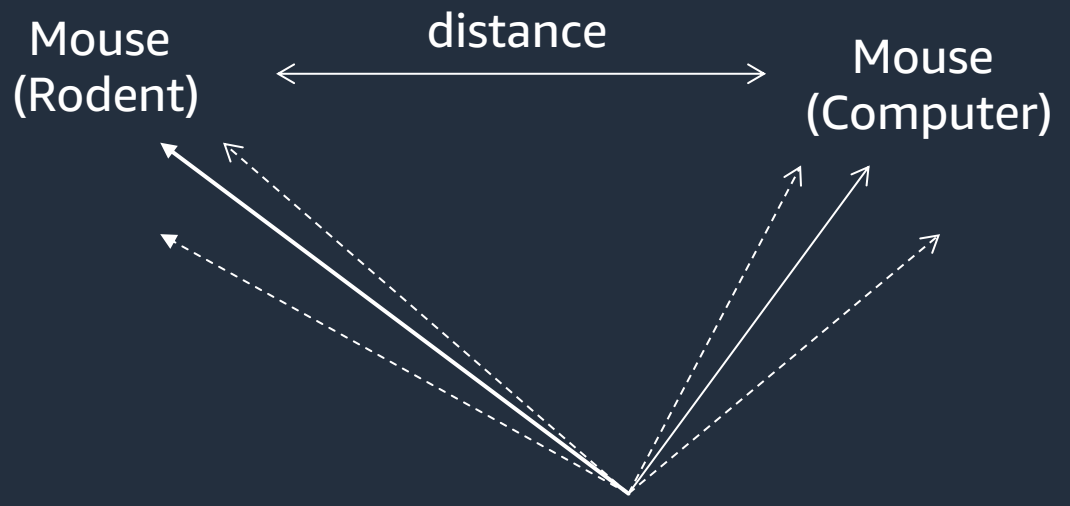


Evaluation methods for LLM systems:



Qualitative comparison and not drawn to scale

Semantic similarity evaluation



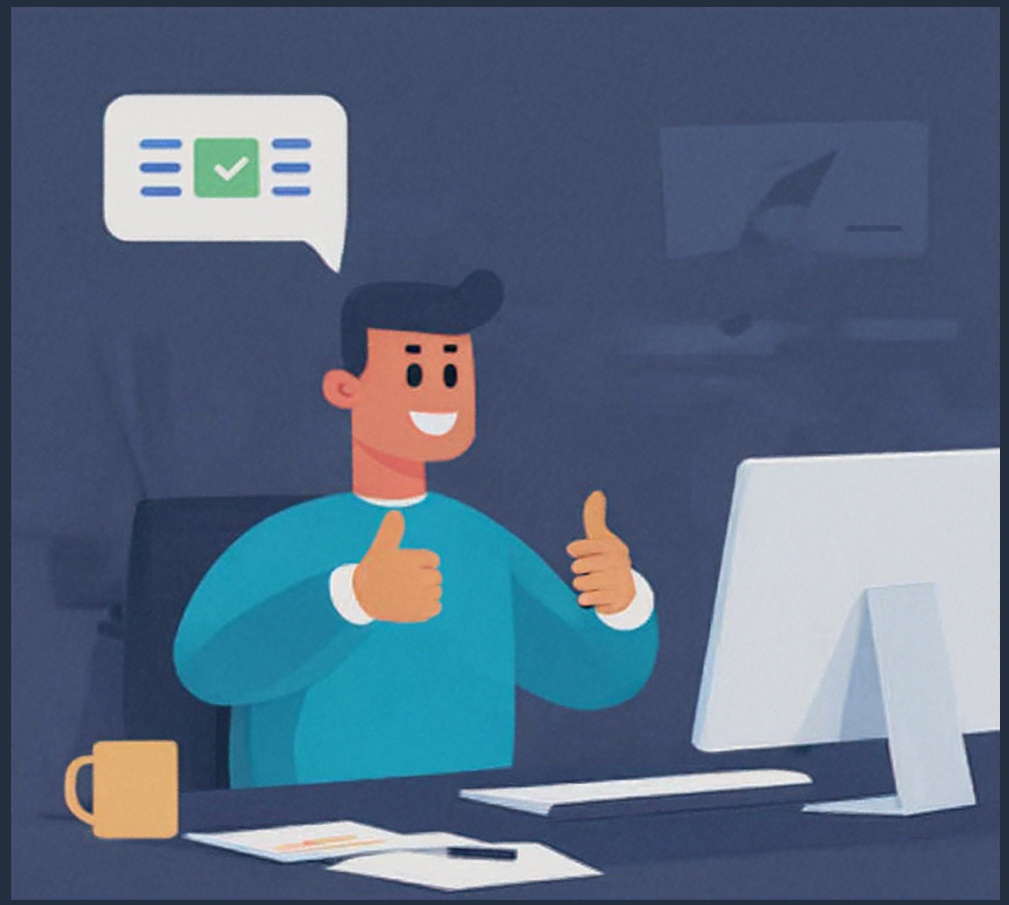
Source: https://medium.com/mlearning-ai/getting-contextualized-word-embeddings-with-bert-20798d8b43a4

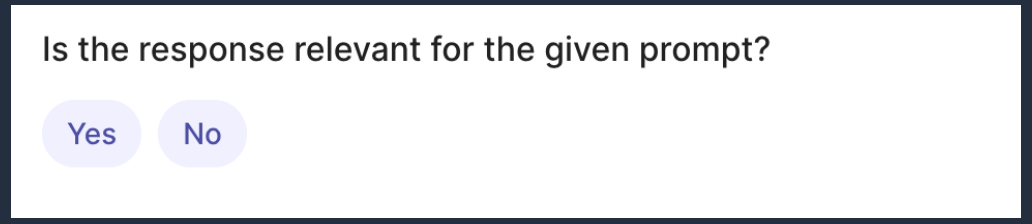
```python
from bert_score import score

with open("hyps.txt") as f:
    cands = [line.strip() for line in f]

with open("refs.txt") as f:
    refs = [line.strip() for line in f]

P, R, F1 = score(cands, refs, lang='en', verbose=True)
```

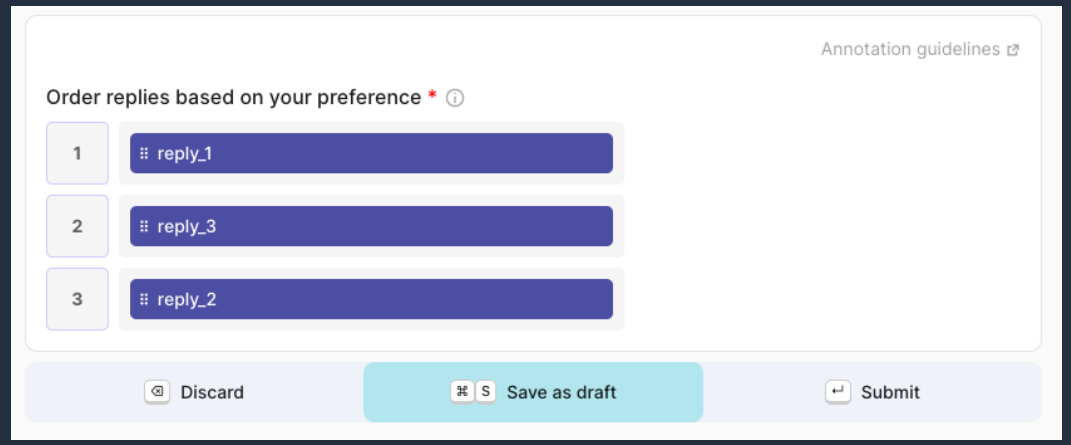
Evaluation methods for LLM systems:

Qualitative comparison and not drawn to scale

Human in the loop / human evaluation







Source: https://gitub.com/argilla-io/argilla

Instructions

Task description

In this task, you will review several responses to a prompt and then rank them in order of clarity.

Instructions

1. Read the provided prompt
2. Rank the three responses on a scale of least to most clear using the tools in the right panel. Each response must be a unique ranking.

Prompt

Explain why you can see myself in a mirror at a level that a 16 year old person can understand

Response A

When light is emitted from a light source like a light bulb, some of it travels toward your body, where it may be reflected toward the mirror with some probability or it may be absorbed. If it were reflected off your body, then some of it could travel toward the mirror, where it could be reflected again. If it is the case where light strikes the mirror, the light is then again redirected as a reflection. If that light is pointed in the direction of your eyes, then the light will enter your eyes. Then, your brain processes the electrical signal made by your eyes and sees it as an image.

Response B

You can see yourself because of a series of light reflections. Light travels from the light source, hits and reflects off of your body and travels toward the mirror. Then, it reflects off of the mirror and travels to your eyes, where your brain registers it as a picture of yourself.

Response C

Light travels in various directions from a light source like a light bulb. Some of the light reflects off of your body with some probability, after which some of it travels to the mirror. Upon striking the mirror, the some of the light again reflects off the mirror and travels toward your eyes, wherein your eyes detect the light after absorbing it. After this process, your brain processes the signal as an image of yourself.

Rank responses by clarity

Response A clarity

Least | Somewhat | Most

Response B clarity

Least | Somewhat | **Most**

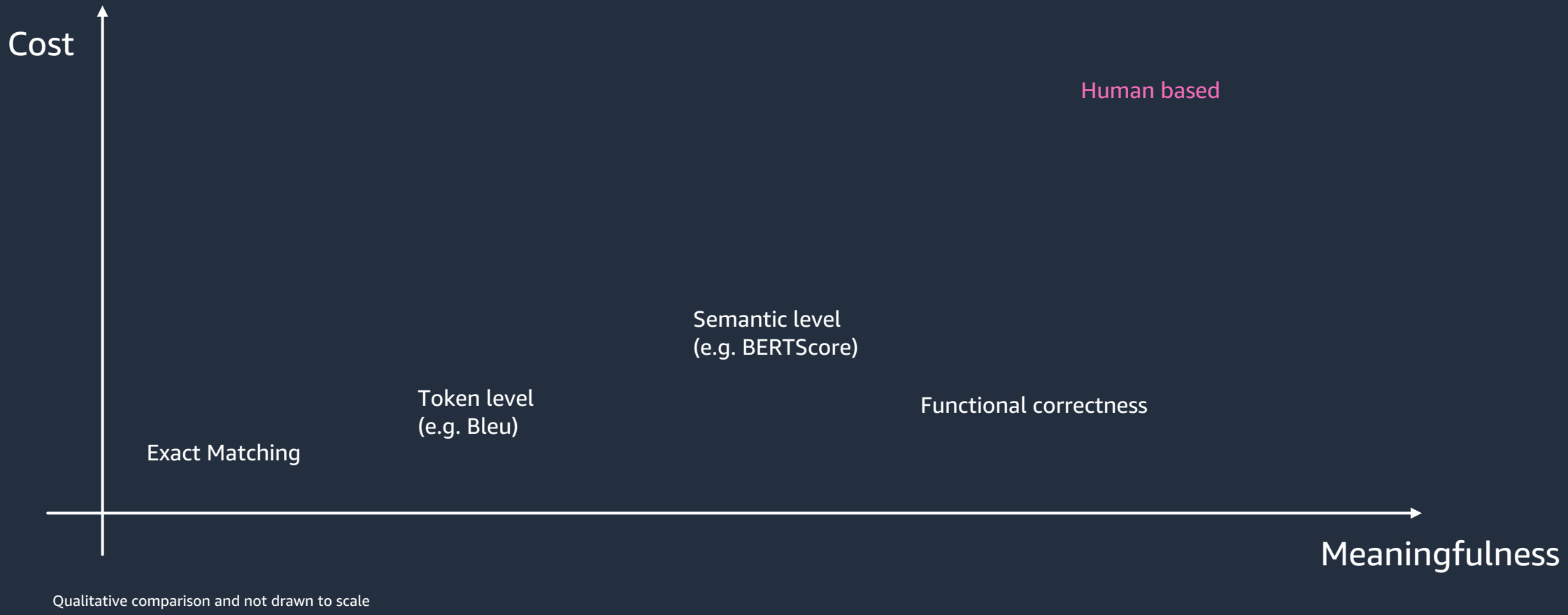
Response C clarity

Least | **Somewhat** | Most

Comments

Save Submit

Evaluation methods for LLM systems:



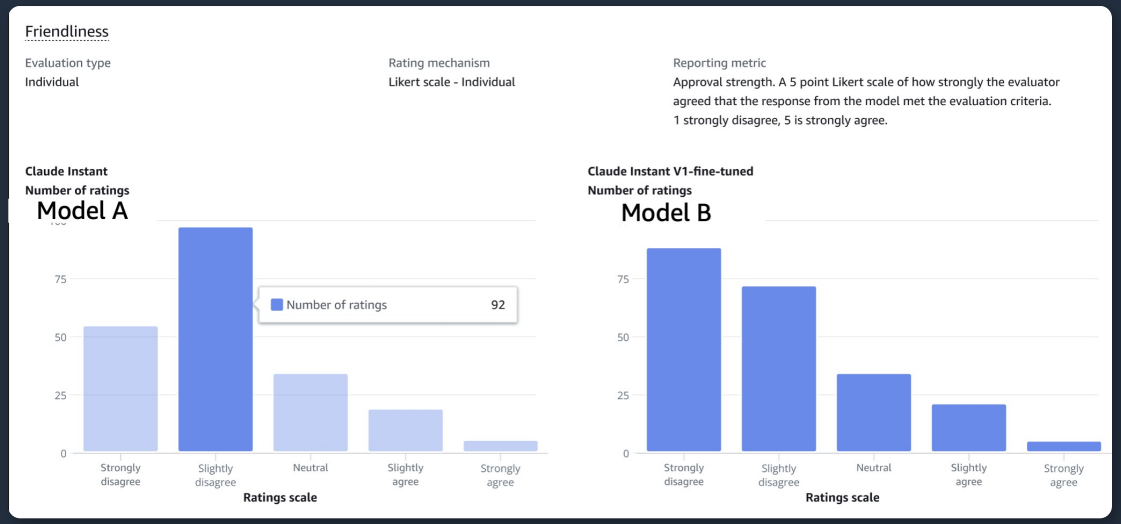
Qualitative comparison and not drawn to scale

Bedrock Evaluation

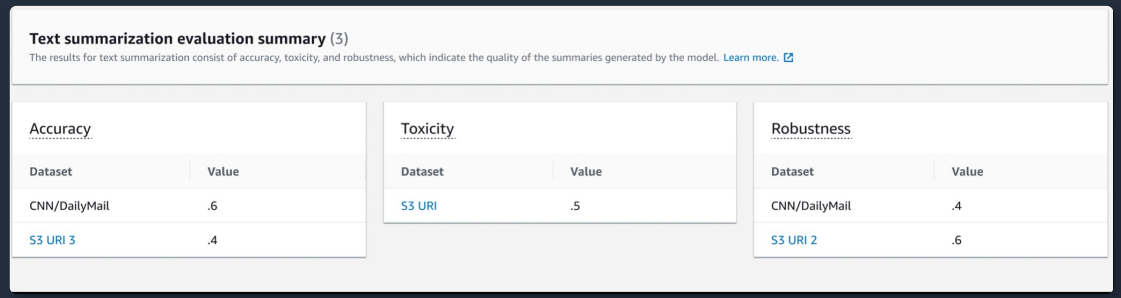
Preview

- Choose automatic or human evaluation method
- Curated datasets or bring your own
- Pre-defined and custom metrics

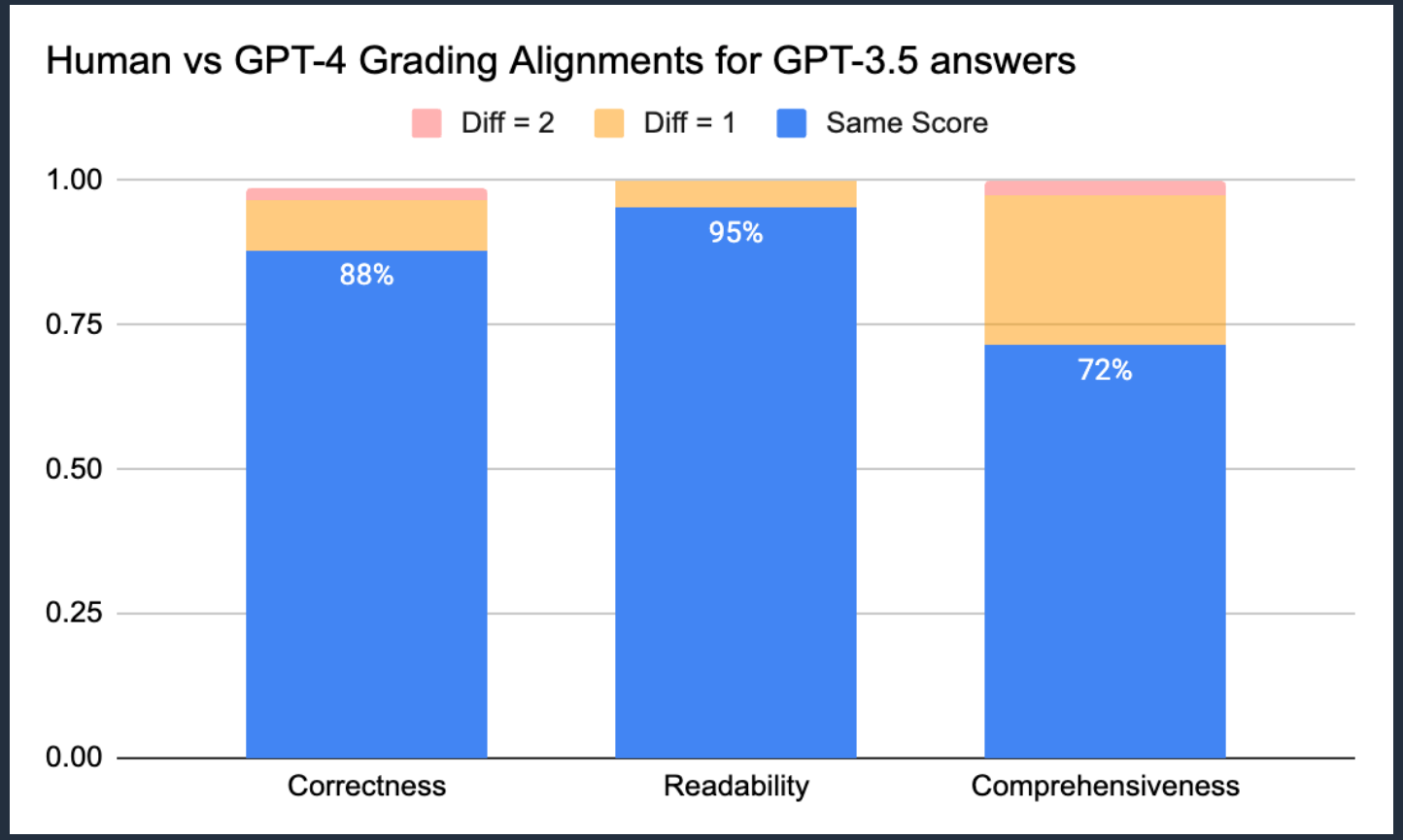
Human evaluation report



Automatic evaluation report



Comparing Human & LLM based evaluation



Source: https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG

Implementing LLMs as a judge

Amazon Bedrock

+

PROMPT

ASK

DEFINITION

GRADING

Create an LLMs as a judge

Amazon Bedrock

+

PROMPT

ASK

DEFINITION

GRADING

You are a language analyst asked to provide feedback to a generated text based on the metric professionalism [...]

Source: https://www.databricks.com/blog/announcing-mlflow-28-llm-judge-metrics-and-best-practices-llm-evaluation-rag-applications-part

Create an LLMs as a judge

Amazon Bedrock

+

PROMPT

ASK

DEFINITION

GRADING

Professionalism refers to the use of a formal, respectful, and appropriate style of communication that is tailored to the context and audience. It often involves avoiding overly casual language [...]

Source: https://www.databricks.com/blog/announcing-mlflow-28-llm-judge-metrics-and-best-practices-llm-evaluation-rag-applications-part

Create an LLMs as a judge

Amazon Bedrock

+

PROMPT

ASK

DEFINITION

GRADING

Score 1: Language is extremely casual, informal, and may include slang or colloquialisms. Not suitable for professional contexts.[...]
Score 5: Language is noticeably formal, respectful, and avoids casual elements. Appropriate for formal business or academic settings.

Source: https://www.databricks.com/blog/announcing-mlflow-28-llm-judge-metrics-and-best-practices-llm-evaluation-rag-applications-part

Let the LLM decide

Amazon Bedrock

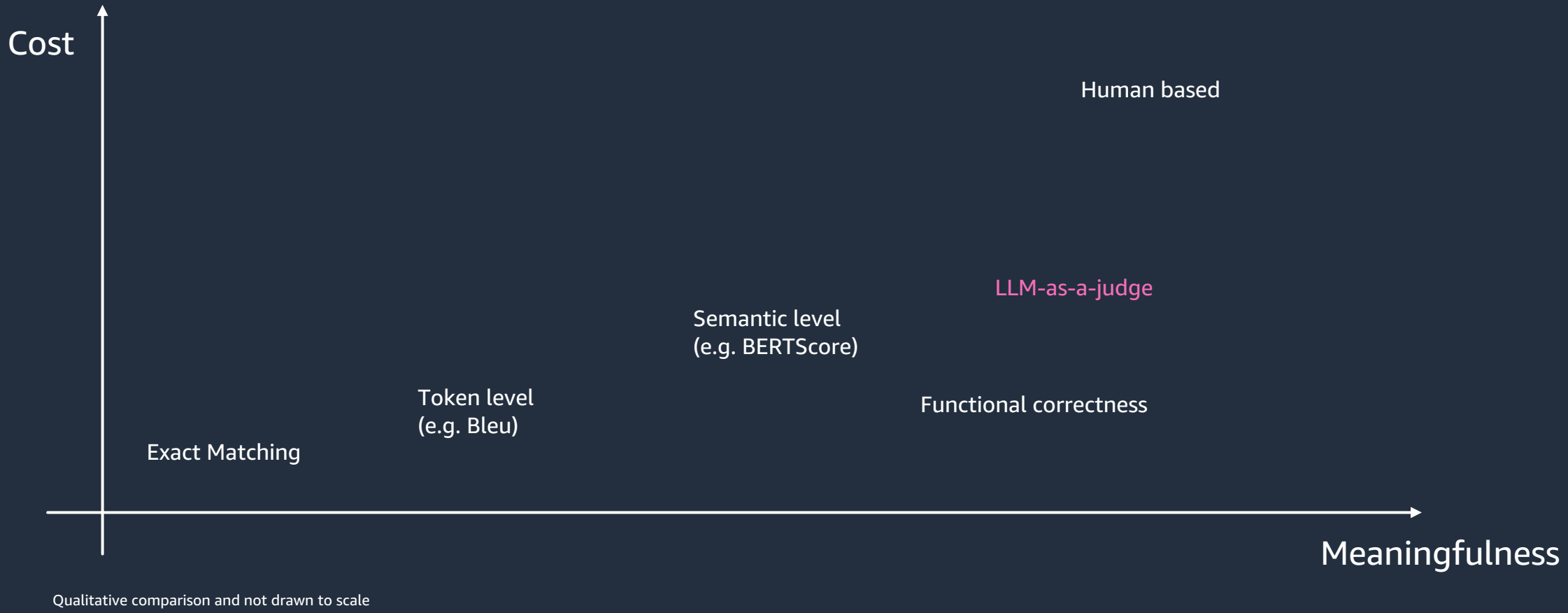
“Based on the metric <professionalism> rank the 3 generated <completions>.[…]”

Be aware of Biases

Bias	Mitigation
Position bias	Switching position of answers
Verbosity bias	Similar completion length
Self enhancement bias	Use different LLMs for generation and validation

Source: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

Evaluation methods for LLM systems:



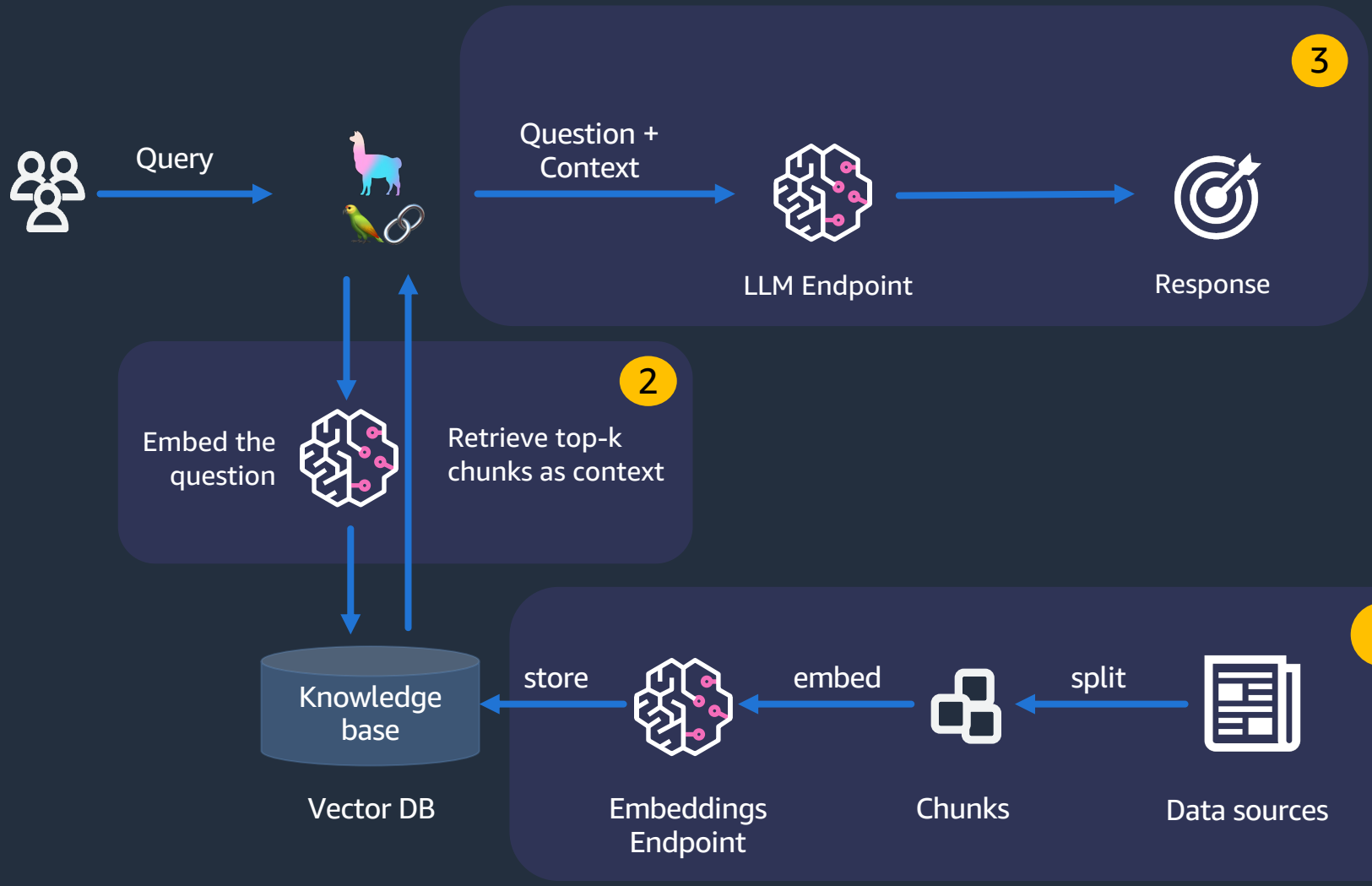
Qualitative comparison and not drawn to scale

Key takeaways

- Traditional NLP metrics don’t align well with human preferences for complex tasks
- Choose the optimal metric and evaluation method based on the task and resources at hand
- Combine multiple evaluation method types to find a good balance between cost, speed and accuracy

Evaluating Retrieval Augmented Generation (RAG)

Recap RAG Architecture



1 Ingestion

2 Retrieval

3 Generation

Evaluating your RAG system

Retrieval

Generation

Evaluating retrieval

- You can use traditional information retrieval metrics like
  - Recall@K - How many relevant contexts did the system retrieve?
  - Precision@K – How many of the retrieved contexts are relevant?
  - Mean Reciprocal Rank (MRR) – How high up is the first relevant retrieved context?
  - Mean average Precision@K (MAP) – How high up are all the relevant retrieved context?

Evaluating generation

Query

Answer Relevance
Is the answer relevant to the query?

Context Relevance
Is the retrieved context relevant to the query?
(Recall, MRR, MAP)

Response

Groundedness
Is the response supported by the context?

Context

https://www.trulens.org/trulens_eval/core_concepts_rag_triad/

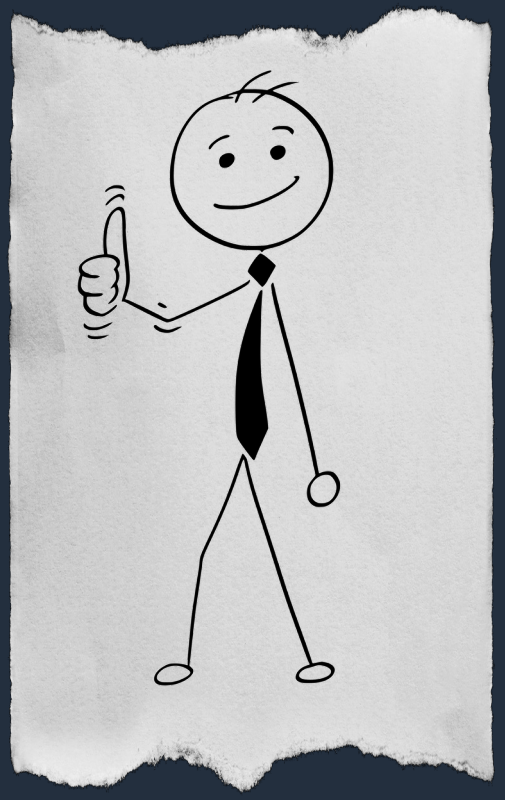
Simplified Prompt – Answer Relevance

<role> You are a RELEVANCE grader; providing the relevance of the given RESPONSE to the given PROMPT. </role>
<task> Respond only as a number from 0 to 10 where 0 is the least relevant and 10 is the most relevant. </task>
<rules>GRADING INSTRUCTIONS: {instructions} </rules>
<input>
PROMPT: {prompt}
RESPONSE: {response}
</input>
RELEVANCE:

https://github.com/truera/trulens/blob/main/trulens_eval/trulens_eval/feedback/v2/feedback.py

Key takeaways – Evaluating RAG

- implement evaluations for each of the components (Retrieval, Generation)
- automate evaluations as much as possible, e.g. using LLM-as-a-judge, augment with human evals (expensive, hard to scale)
- join the next session “Ensuring accuracy of LLM responses with state-of-the-art Information Retrieval“ to dive deeper into RAG evaluation

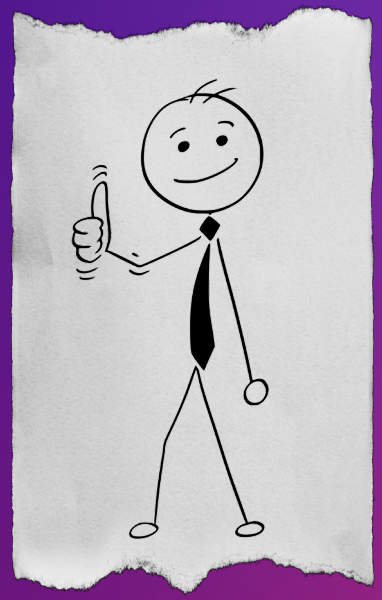


- Invest in building a test set and automate evaluation from the beginning
- Pick the right metric and evaluation method based on your task and resources
- Use your evals to guide you on what works and what doesn’t

Thank you!

Johannes Langer
johlanger

Lukas Wenzel
lukwenzel



Please complete the session survey.