



# QUESTION ANSWERING

## Information Retrieval System

DATA SCIENCE INTENSIVE CAPSTONE  
BY: DENNIS DANG

# Our Client

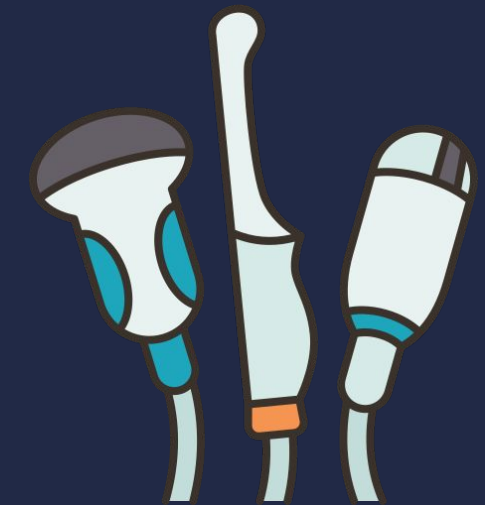
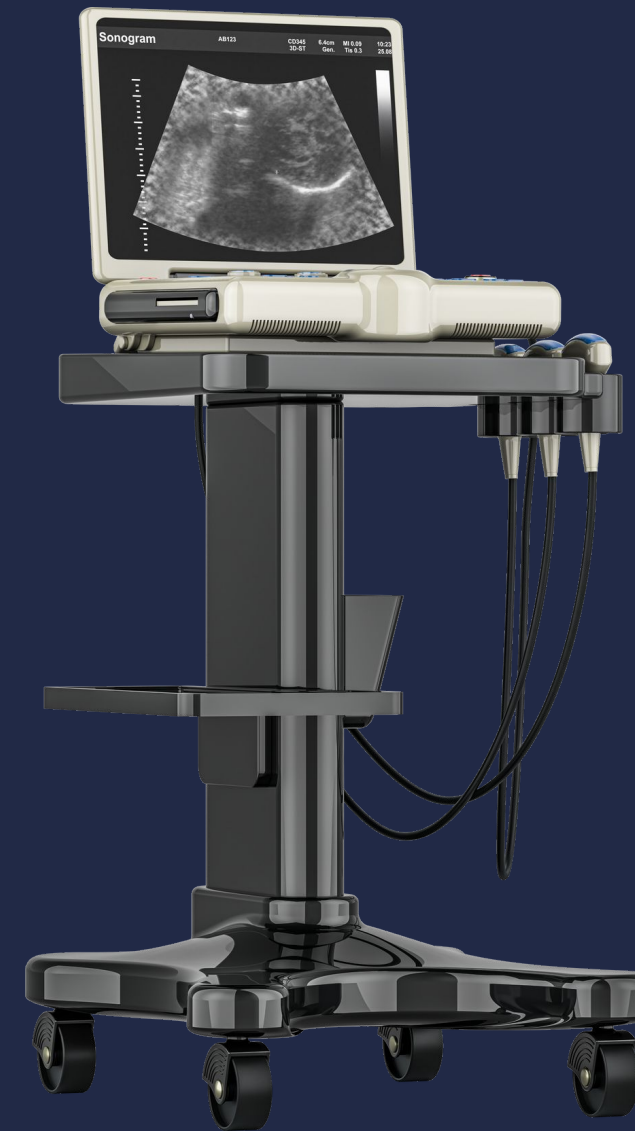
2D Imaging, Inc. is a small, family-owned business specializing in medical ultrasound equipment sales, repairs, refurbishing, and servicing.

The company is currently expanding its e-commerce capabilities and online sales presence.

- Highly specialized, expensive medical equipment **requires** good customer service
- Customer support infrastructure for increased volume and diversity of inquiries

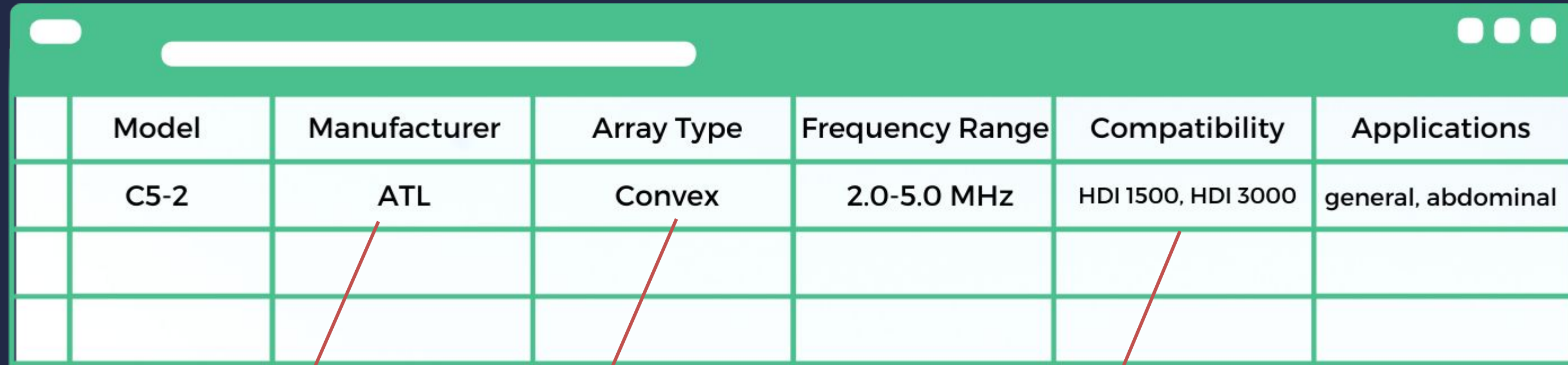
## Goal

- Develop a Question Answering (QA) system to provide accurate, timely, and relevant information to customers



# Preparing the knowledgebase

- Loop through tabular data and transcribe into human-readable sentences



	Model	Manufacturer	Array Type	Frequency Range	Compatibility	Applications
	C5-2	ATL	Convex	2.0-5.0 MHz	HDI 1500, HDI 3000	general, abdominal

*"The manufacturer of the C5-2 probe is ATL."*

*"The ATL C5-2 is a convex array type probe."*

*"The ATL C5-2 probe is compatible with the following systems: HDI 1500, HDI 3000."*

**Final corpus contains 324 text documents**



# Generating Question-Document Pairs

- Used large language model (Llama3) to generate 3 questions for each document

```
Generate three concise questions that can be answered using the following information, similar to the example provided. Provide only the questions, numbered as follows:
```

```
1. [insert question here]
2. [insert question here]
3. [insert question here]
```

Example 1:

Information: The manufacturer of the C3 probe is ATL.

Questions:

```
1. Who is the manufacturer of the C3 probe?
2. Who makes the C3 transducer?
3. Is the C3 probe made by ATL?
```

Example 2:

Information: The G.E. RIC5-9D probe is compatible with the following systems: Voluson E6, Voluson E8, Voluson E10, LOGIQ S7, LOGIQ S8, Vivid E95, LOGIQ E9, LOGIQ E10.

Questions:

```
1. What systems are compatible with the G.E. RIC5-9D probe?
2. Does the G.E. RIC5-9D work with the Voluson E6 system?
3. Is the G.E. RIC5-9D transducer compatible with the LOGIQ S7 system?
```

Example 3:

Information: The Siemens Acuson 15L8W probe has a variant with a cartridge connection.

Questions:

```
1. Does the Siemens Acuson 15L8W probe have a cartridge connection?
2. What kind of connector does the Siemens Acuson 15L8W transducer use?
3. Does the Siemens Acuson 15L8W probe use a cartridge connector?
```

## Prompt Engineering

- Provide examples for each type of document
- Provide clear instructions for parseable output structure (e.g. numbered list)

```
Here are the three questions generated for the documents:
```

```
1. Who is the manufacturer of the C3 probe?
2. Who makes the C3 transducer?
3. Is the C3 probe made by ATL?
```

```
def extract_questions(response):
    pattern = r'\d+\.\s(.*?)'
    questions = re.findall(pattern, response)
    return questions
```

```
["Who is the manufacturer of the C3 probe?", "Who makes the C3 transducer?", "Is the C3 probe made by ATL?"]
```

# Generating Question-Document Pairs

- Used llama3 to filter out poorly generated question-document pairs
- Any question-document pairs marked as irrelevant were manually reviewed
- Final dataset: **932 queries, 324 documents**

```
grading_prompt_template = PromptTemplate(  
    input_variables=["content", "question"],  
    template="""  
    Given the content: "{content}"  
    Can the following question be sufficiently answered?  
    Question: "{question}"  
    Your response should only consist of one number, either a 0 (meaning No) or 1 (meaning Yes).  
  
    Example 1:  
    Content: "The manufacturer of the C3 probe is ATL."  
    Question: "Who is the manufacturer of the C3 probe?"  
    Response: 1  
  
    Example 2:  
    Content: "Who manufactures the EPIQ 7 ultrasound system?"  
    Question: "The Siemens Acuson 12L3 probe is compatible with the following systems: Juniper."  
    Response: 0  
    """)
```



# Evaluating Pre-trained Models

## all-mpnet-base-v2

- General-purpose model based on the MPNet architecture
- Trained on diverse corpus of 1B+ sentences
- <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

## multi-qa-mpnet-base-dot-v1

- mpnet-base model fine-tuned using 215M question-answer pairs
- <https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1>

## multi-qa-distilbert-cos-v1

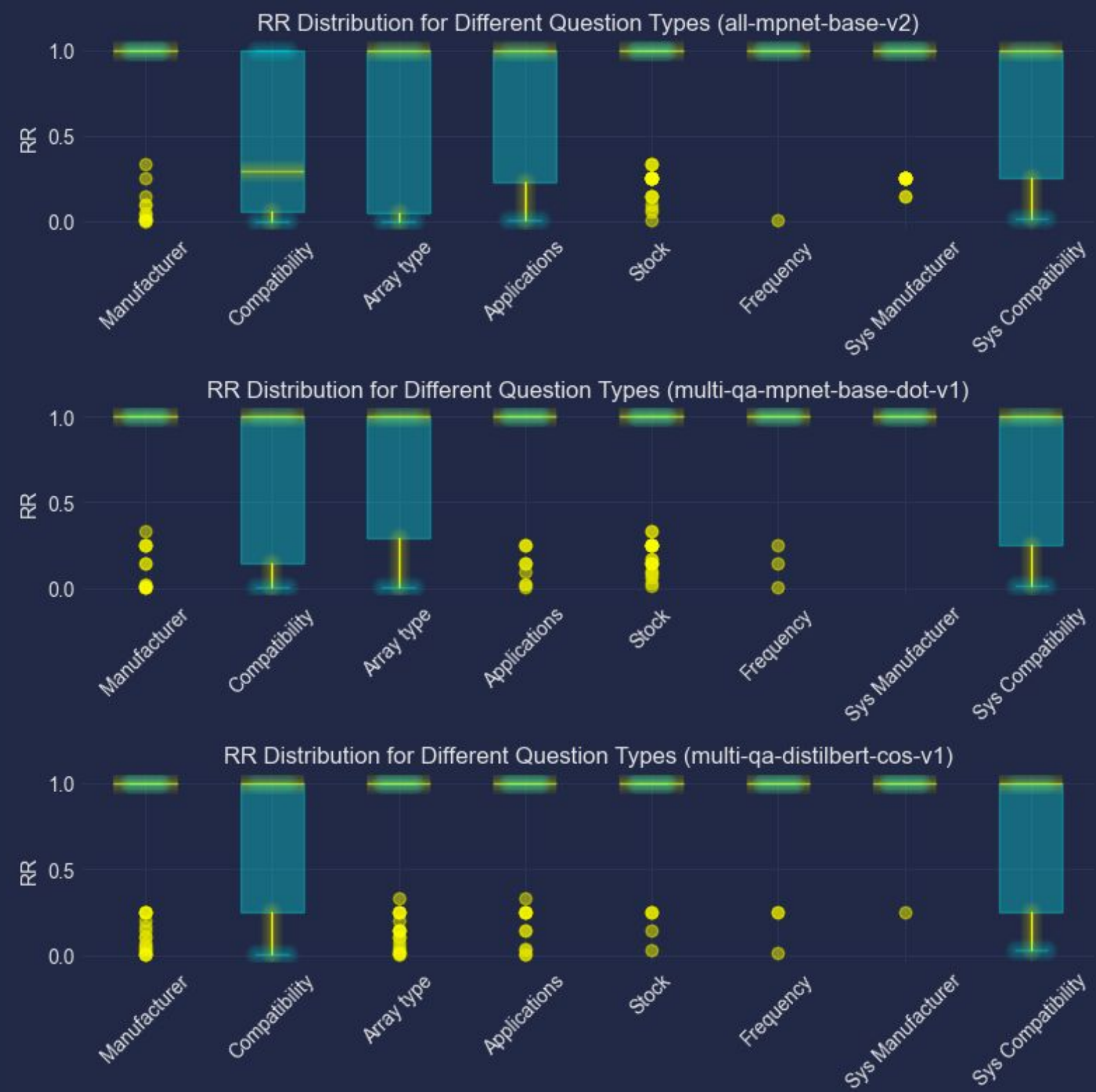
- Fine-tuned using 215M question-answer pairs
- Based on the DistilBERT architecture, which is a lighter and faster version of the popular BERT model
- <https://huggingface.co/sentence-transformers/multi-qa-distilbert-cos-v1>

## Evaluation Metric:

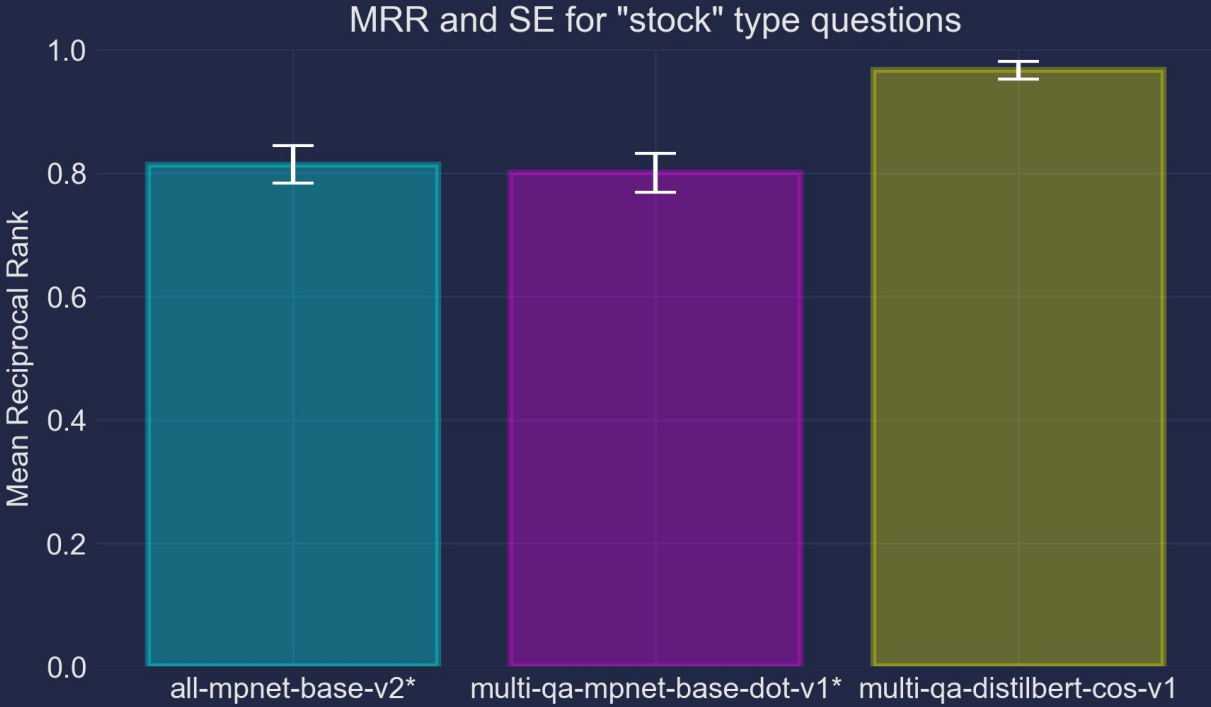
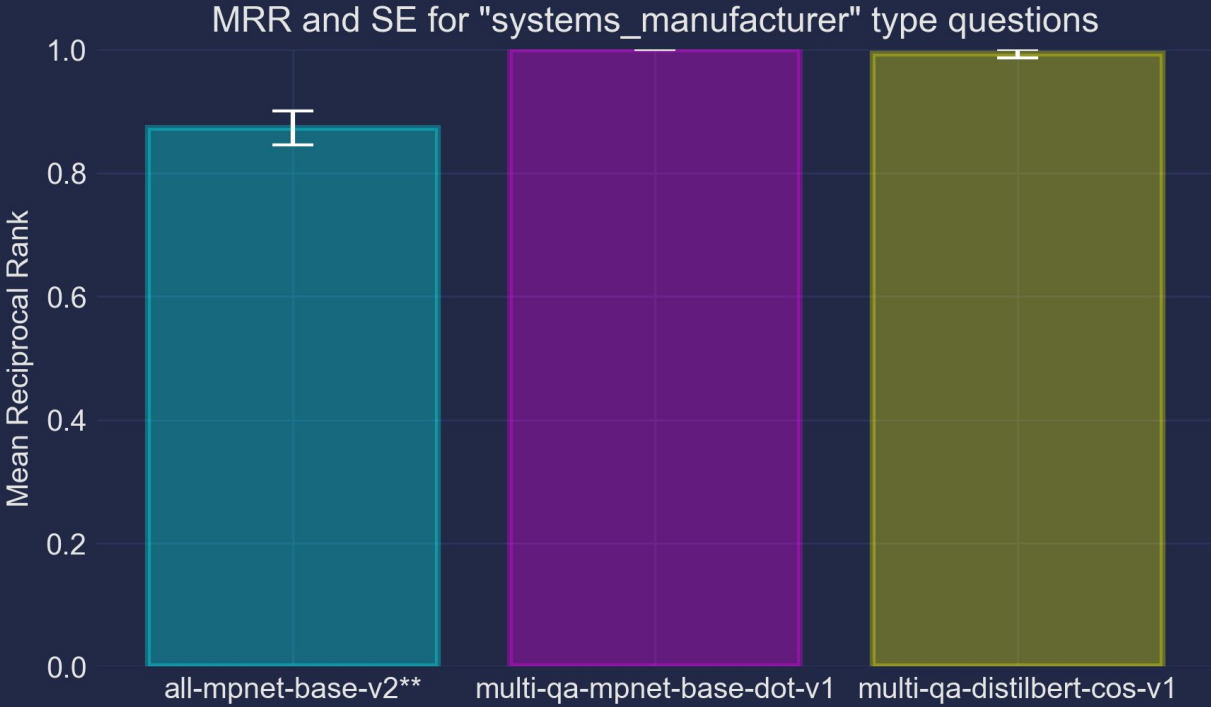
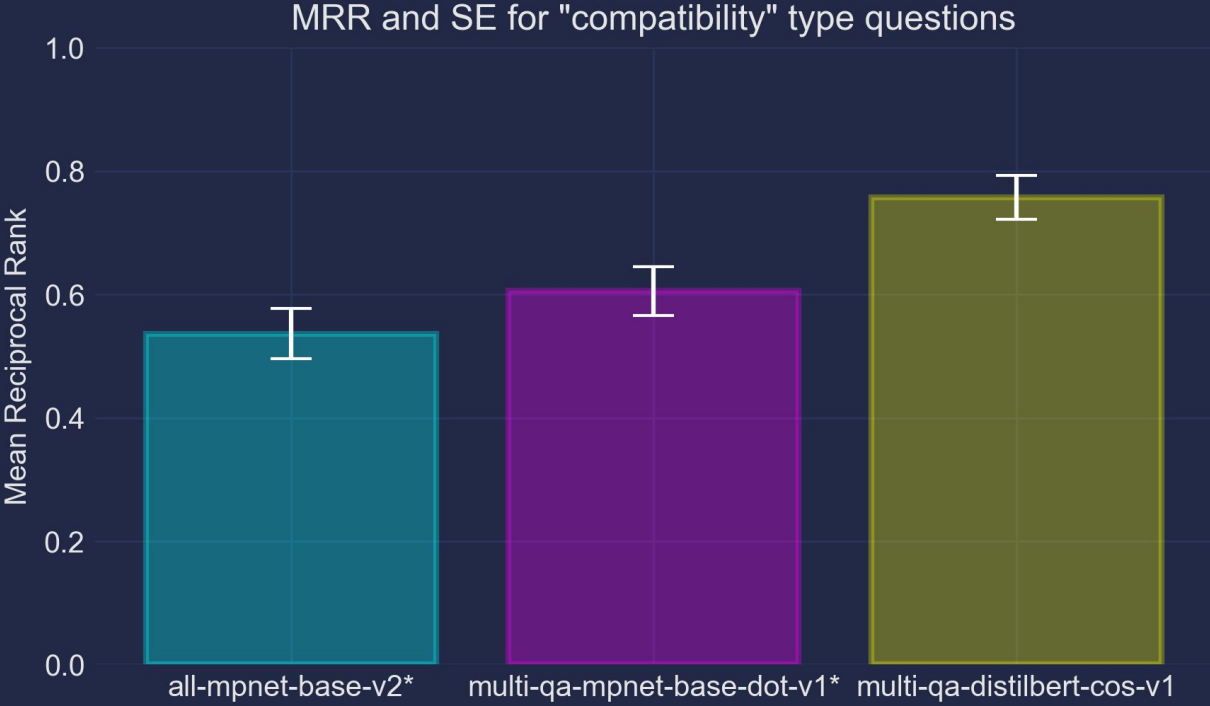
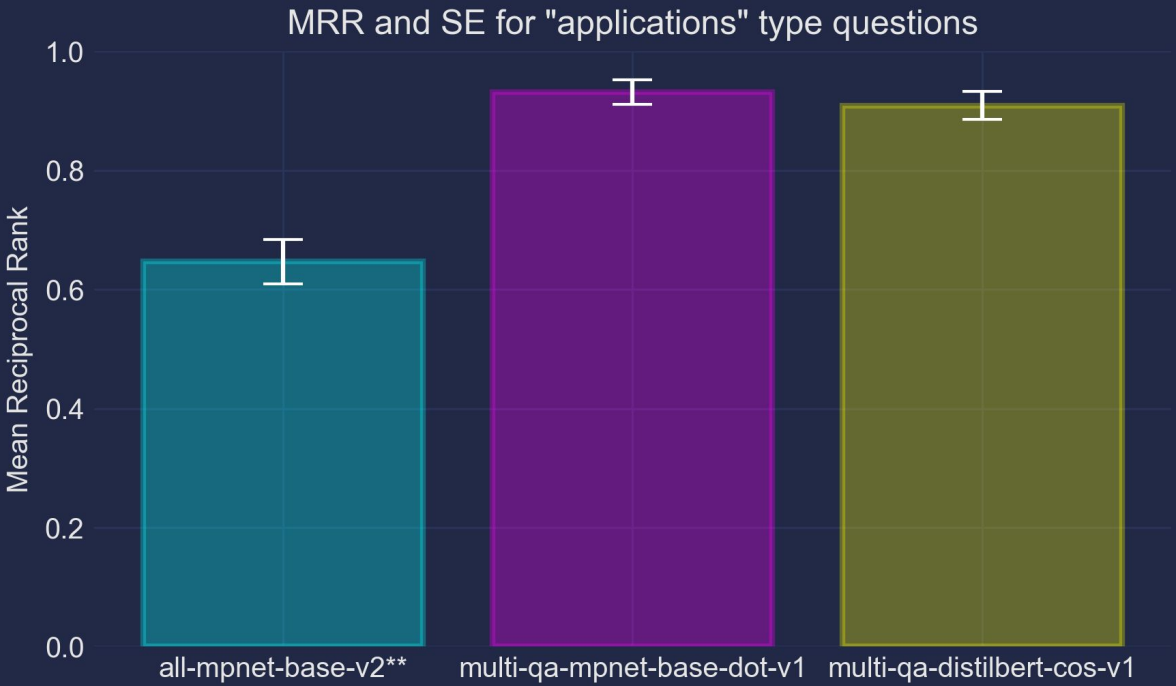
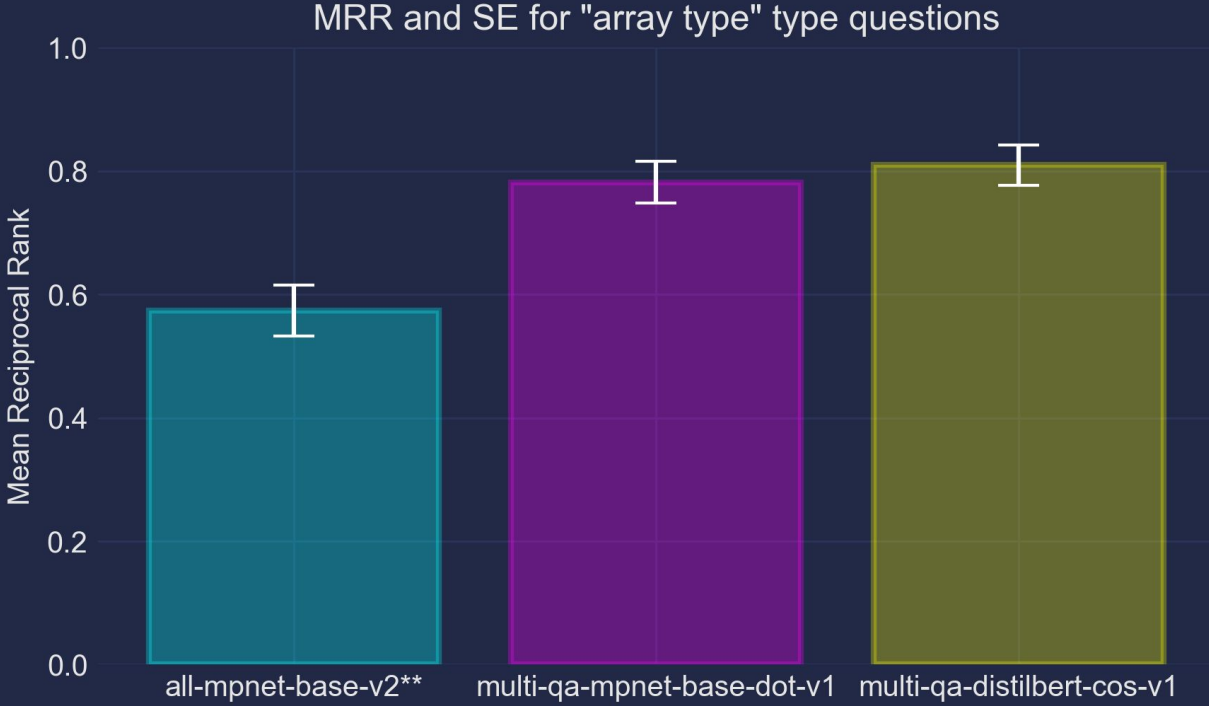
Mean Reciprocal Rank (MRR)

- Documents ranked based on their cosine similarity scores to the query
- Calculates reciprocal rank ( $1/\text{rank}$ ) of the “correct” document

# Distribution of Reciprocal Rank Scores by Question-type and Embedding Model



# Mean Reciprocal Rank and SE across Various Question Categories



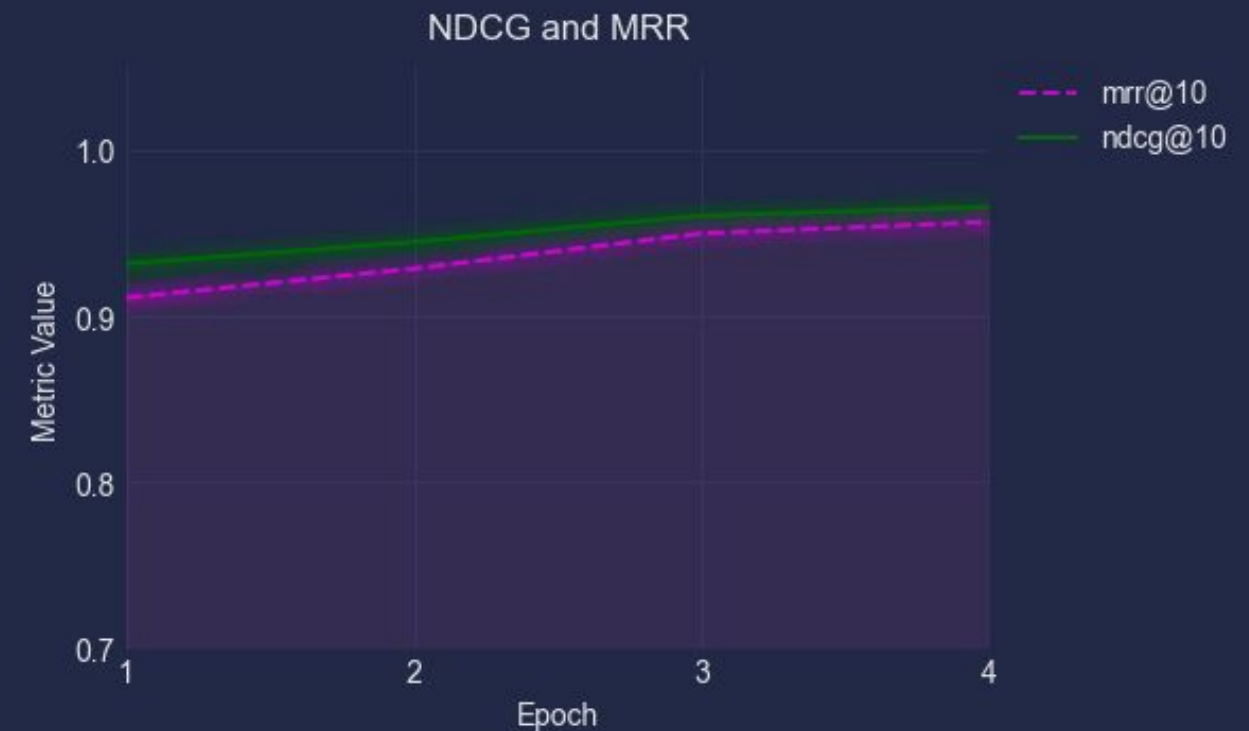


# Hyperparameter Tuning

- **Base model:** *multi-qa-distilbert-cos-v1*
- Create training/validation/testing (70/15/15) PyTorch Datasets
- Bayesian Optimization for hyperparameter tuning
  - per\_gpu\_batch\_size: (16, 64) → 56
  - weight\_decay: (0, 0.3) → 0.21
  - learning\_rate: (1e-5, 5e-5) → 3.6e-5
  - warmup\_steps: (0, 500) → 106
  - num\_epochs: (2, 5) → 4

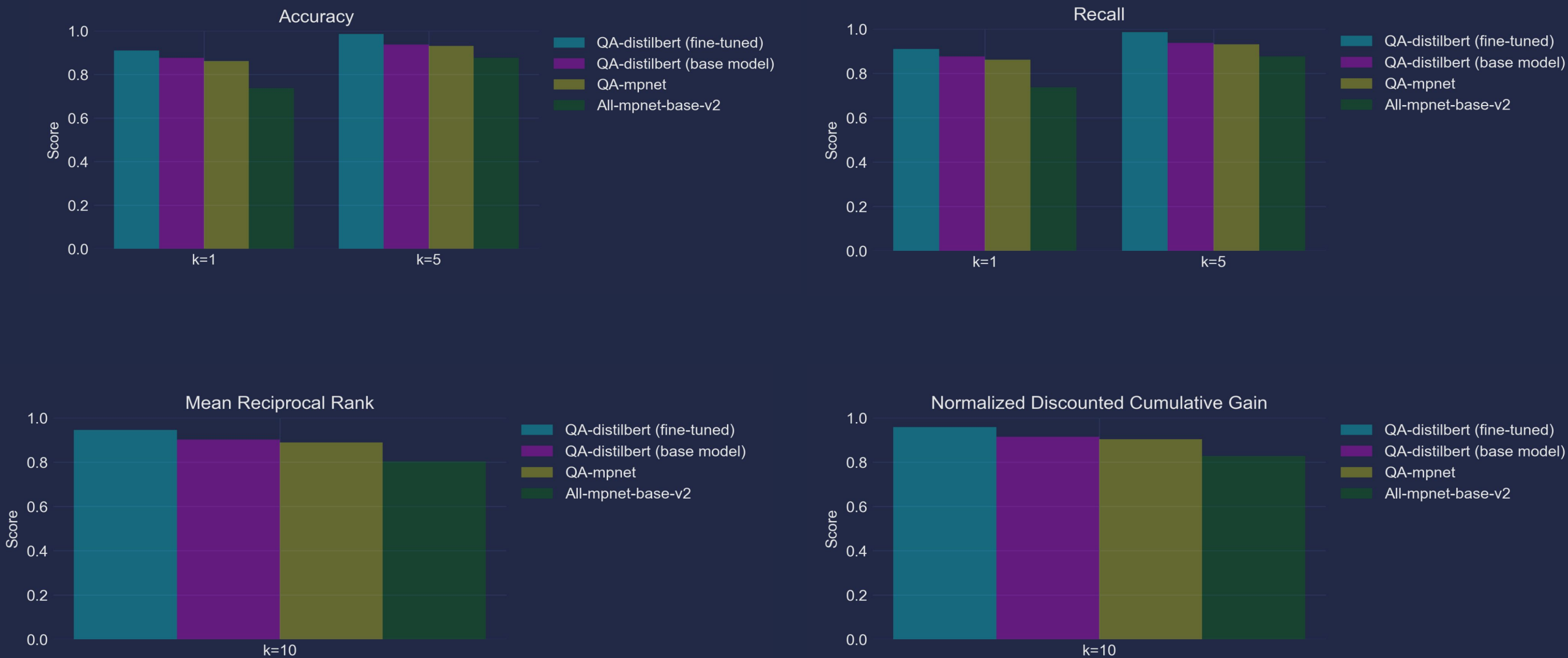
# Fine-tuning (training)

- **Loss function:** *MultipleNegativesRankingLoss*
  - Designed for training models with only positive pairs of data (e.g. query, relevant document)
  - Minimizes negative log-likelihood of softmax-normalized similarity scores
- Performance metrics collected at different top-k values (e.g., @1, @5, @10)
  - Accuracy, Recall, MRR, NDCG (Normalized Discounted Cumulative Gain)



# Model Evaluation and Comparison

Table 1: Performance Metrics for Each Embedding Model				
	QA-distilbert (fine-tuned)	QA-distilbert (base model)	QA-mpnet	All-mpnet-base-v2
Metric				
Accuracy@1	0.9103	0.8759	0.8621	0.7379
Accuracy@5	0.9862	0.9379	0.9310	0.8759
Precision@1	0.9103	0.8759	0.8621	0.7379
Recall@1	0.9103	0.8759	0.8621	0.7379
Recall@5	0.9862	0.9379	0.9310	0.8759
NDCG@10	0.9586	0.9142	0.9027	0.8276
MRR@10	0.9449	0.9020	0.8892	0.8028





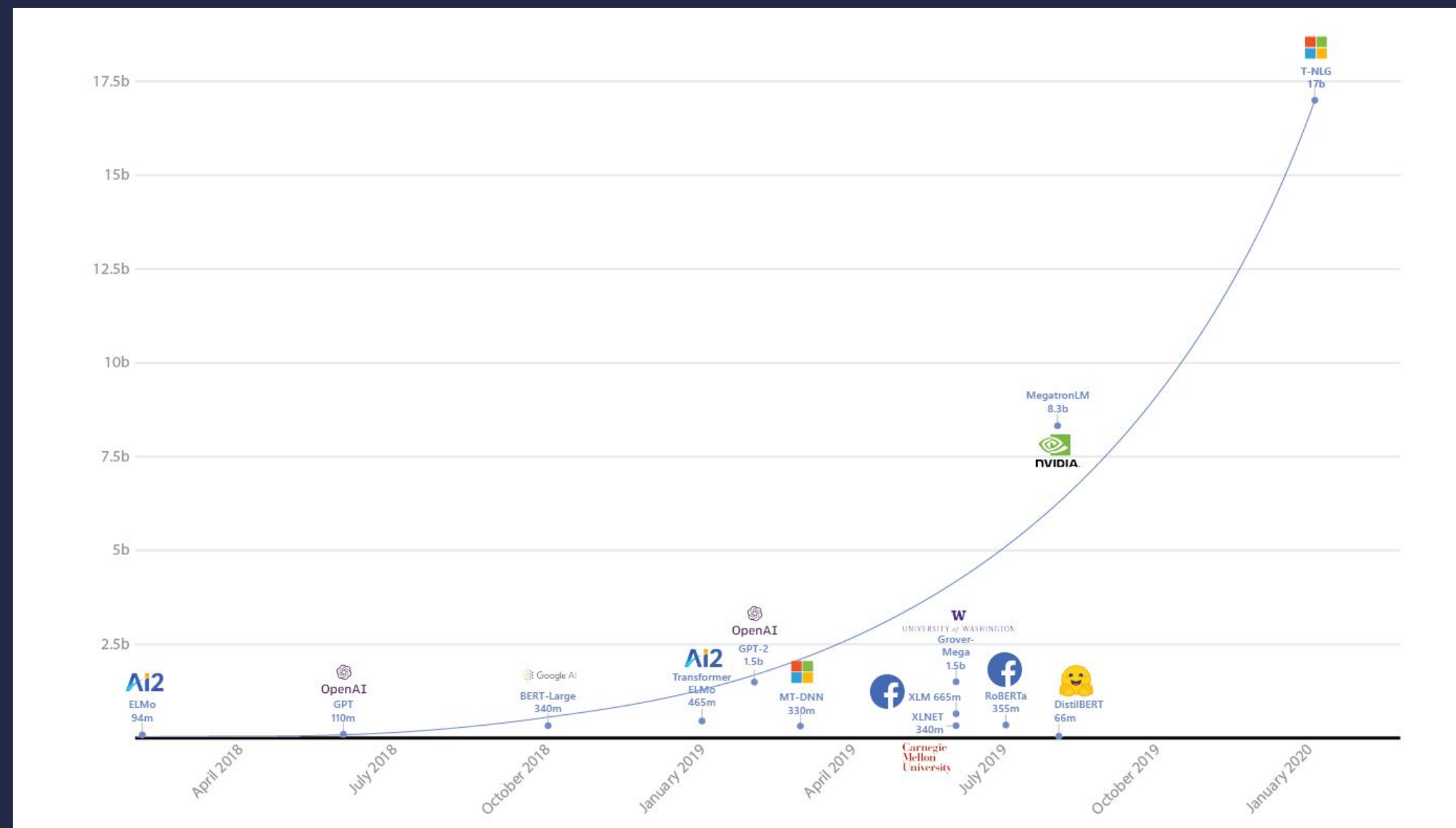
# Summary

- 1** Create the document knowledge base
- 2** Generate queries for each document using LLM
- 3** Evaluate pre-trained embedding models
- 4** Bayesian Optimization
- 5** Fine-tuning
- 6** Evaluate



# Concluding Remarks

- DistilBERT's size makes it a practical choice for customer support systems
- Generated queries are relatively simplistic (required only one document to formulate an answer)
- Expand QA system to retrieve multiple relevant documents and LM to synthesize a response
- Integrate LM chains with structured database queries





**THANK YOU FOR  
LISTENING!**