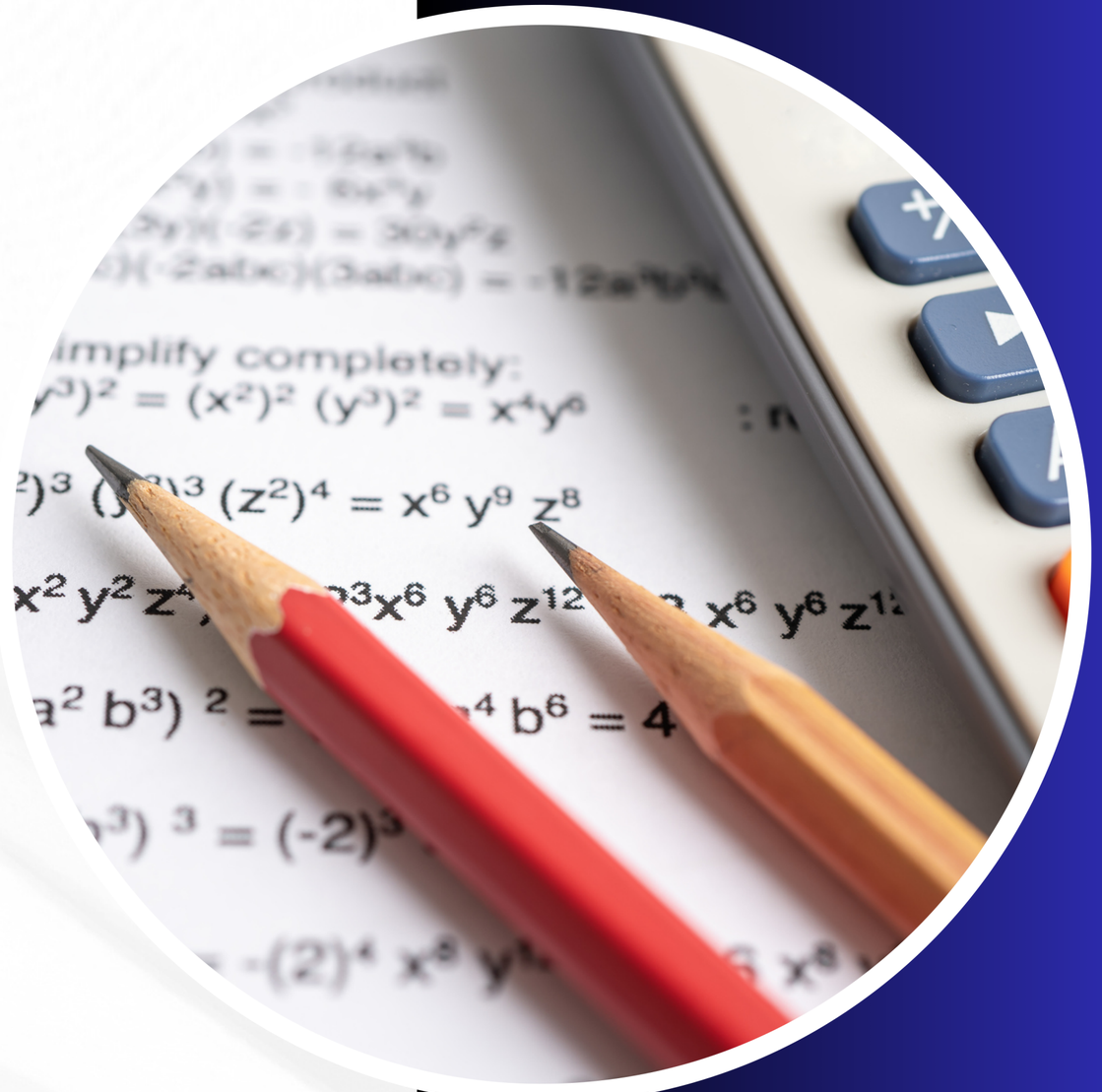
NLP: PROJECT PROPOSAL

MATHPLOT VQA



Presented by:
Sitthiwat Damrongpreechar st123994
Pirunrat Kaewphanoaw st124003
Munthitra Thadthapong st124022
Parun Ngamcharoen st124026
Nathas Sungworawongpana st124323

PROBLEM AND MOTIVATION

Current Question Answering (QA) systems primarily **focus on text-based** inputs, **neglecting** the needs of individuals who struggle with understanding **graphs** in math problems.

Our goal is to develop a VQA system capable of interpreting and responding to questions about graphs, aiming to make math more accessible and understandable for everyone, particularly those facing difficulties with **graph comprehension**

INTERESTING PAPERS

01 Pix2Struct

02 ChartQA

03 MATCHA

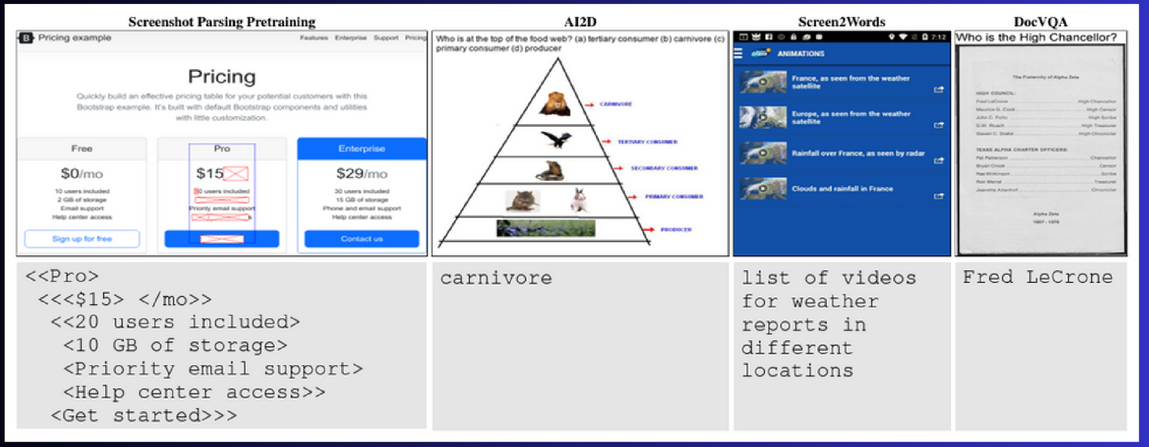
04 LLaVA

PIX2STRUCT

SCREENSHOT PARSING AS PRE-TRAINING FOR VISUAL LANGUAGE UNDERSTANDING

- Introducing Pix2Struct, an image-to-text model pretrained by parsing masked web page screenshots into HTML.

- For fine-tuning, tasks such as common pre-training signals like OCR, language modeling, and image captioning are implemented.

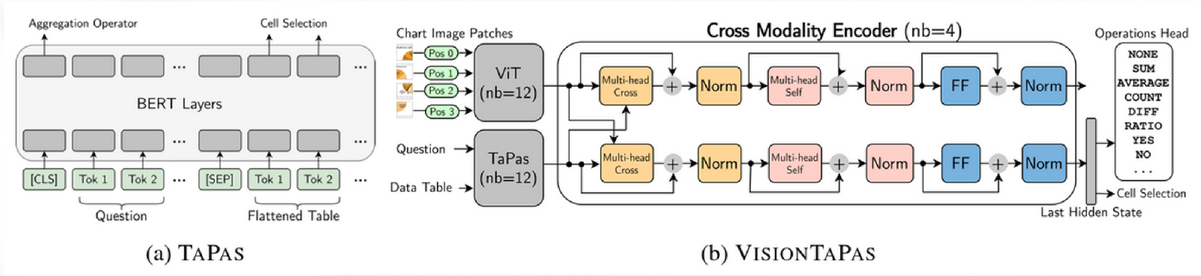


CHARTQA

AI DATA ANALYSIS SOLUTION FOR CHART-BASED QUERIES

- ChartQA introduced benchmark datasets which comprise of human-written questions with chart summaries pairs.

- Moreover, introduced VisionTaPas, which consists of two transformer-based models designed to combine visual features and chart data tables for unified question answering.

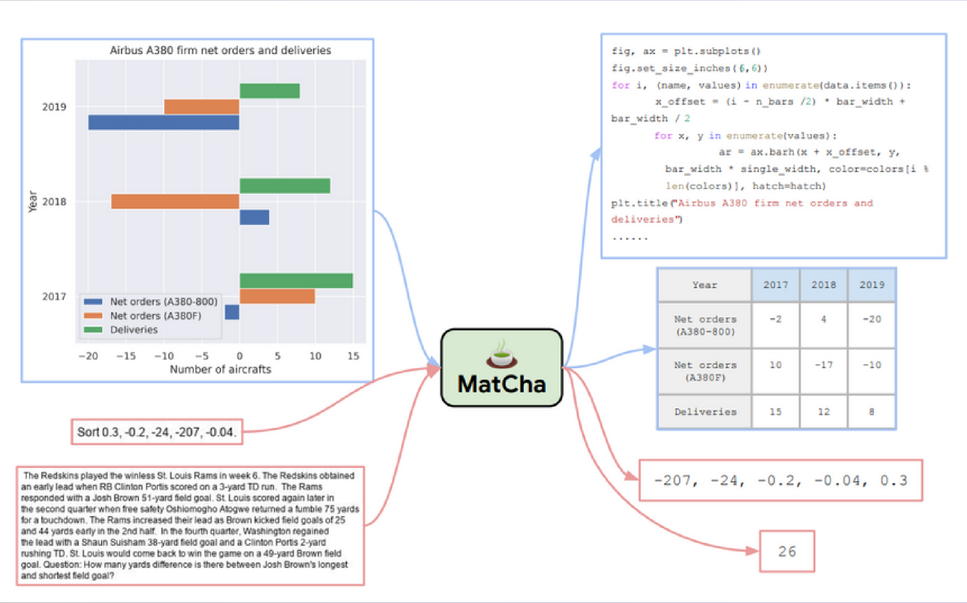


(a) TaPas (b) VisionTaPas

MATCHA

MATH REASONING AND CHART DERENDERING PRETRAINING

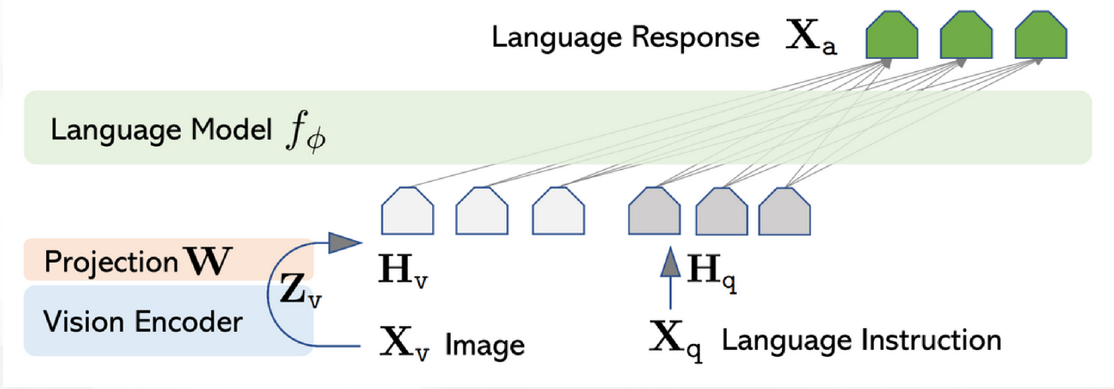
- Enhances visual language models' capabilities in jointly modeling charts/plots and language data.
- MATCHA pretraining starts from **Pix2Struct** as the base model and further pretrain it with chart derendering and math reasoning tasks.



LLAVA

LARGE LANGUAGE AND VISION ASSISTANT

- Training the model on detailed description tasks and complex reasoning tasks also helps boosting the performance of conversational tasks.
- LLaVA uses visual encoder (CLIP) to encode image to text. Then, use the decoded text and query to generate prompt for LLM (Vicuna) training.

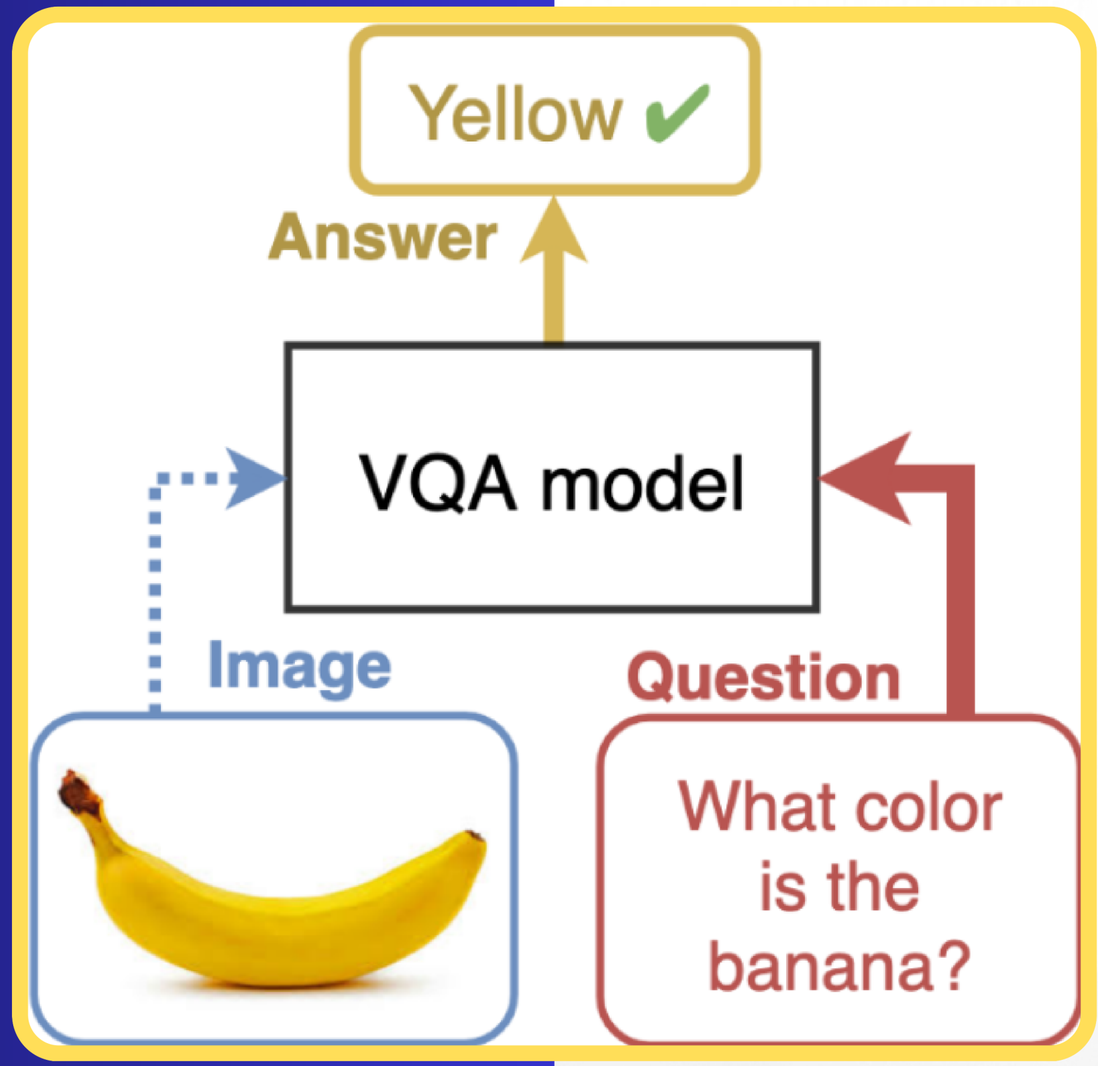


PROPOSED

MATHPLOT VQA

INTRODUCTION
—WHAT IS VQA?

- LLM is a model that accepts text (query) in order to generate texts.
- Visual Question Answering (VQA) is a model that accepts both **image and text** (query) at the same time in order to generate texts**.**
- VQA extends to charts, broadening machine comprehension to include graphical data representations to answer a user-specified query.
- VQA systems aim to bridge the gap between visual understanding and natural language understanding, enabling machines to understand and respond to questions about visual content in a human-like manner.



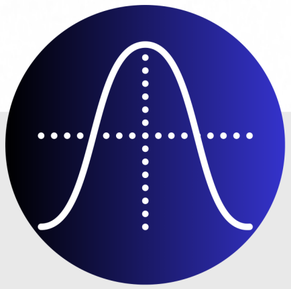
SOLUTION REQUIREMENTS

GRAPH IMAGE SUBMISSION

Users can effortlessly submit mathematical graph images, facilitating seamless interaction.

QUESTION INPUT

The system enables users to ask contextually relevant questions about the submitted graph, fostering active engagement.



MATH QUESTION-SOLVING (FROM GRAPH)

The model provides the solutions to mathematical problems derived from the graph data, along with step-by-step explanations.

GRAPH INTERPRETATION

The system offers insightful analysis of graph data, elucidating key trends and relationships to aid user interpretation.

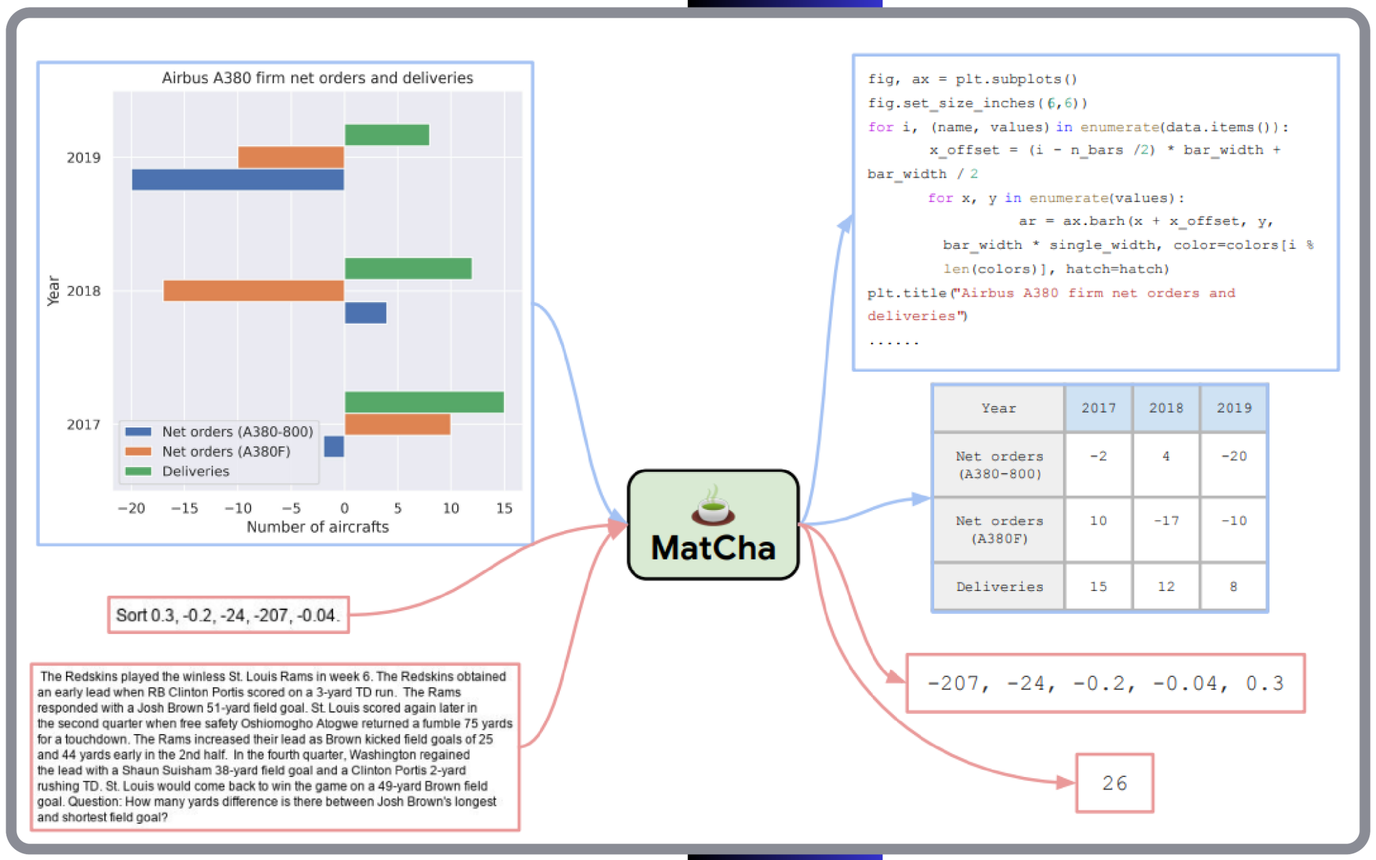
DATASET AND PRETRAINED MODEL

Dataset

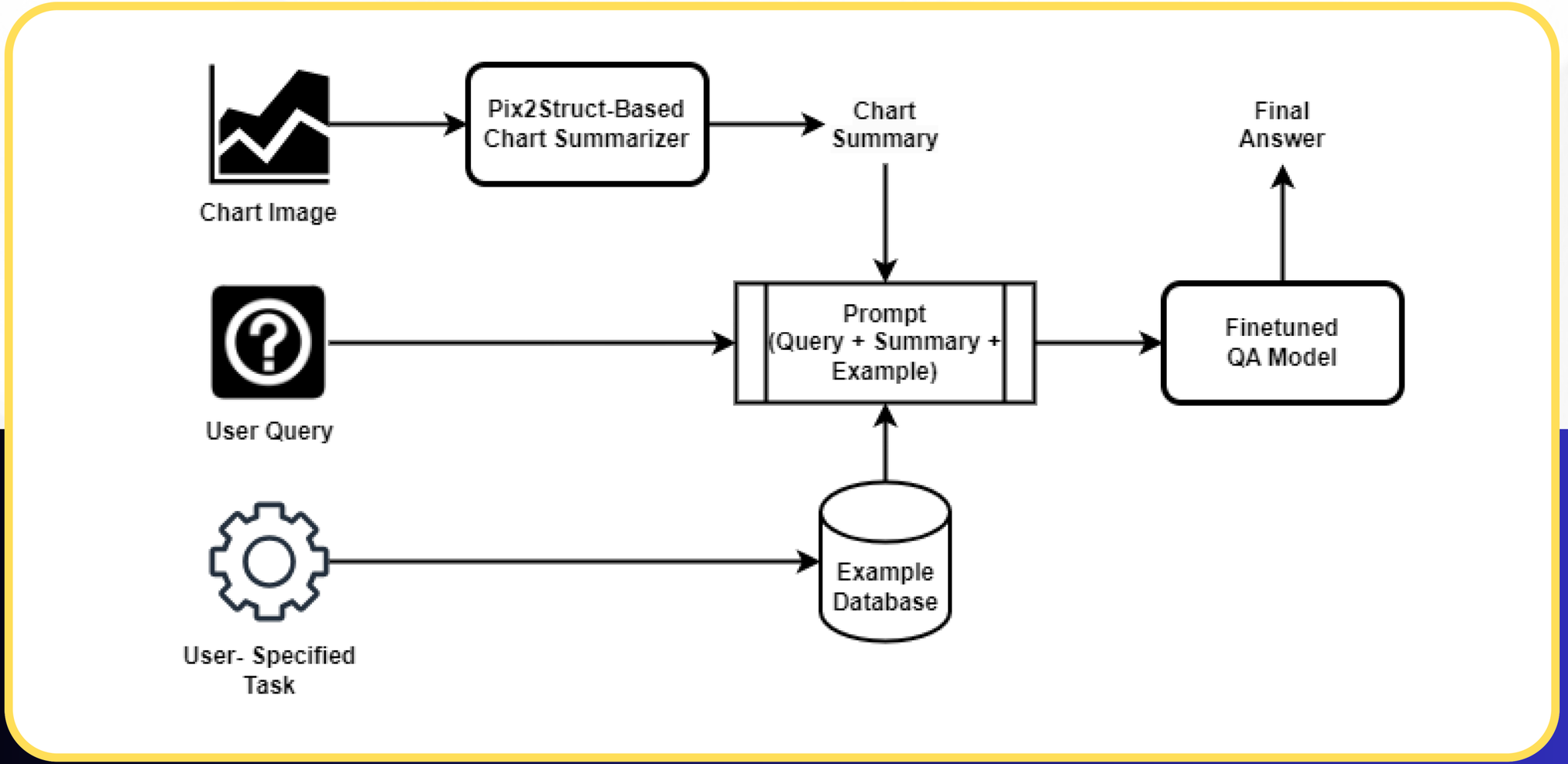
- ChartQA: from github

Pretrained Model

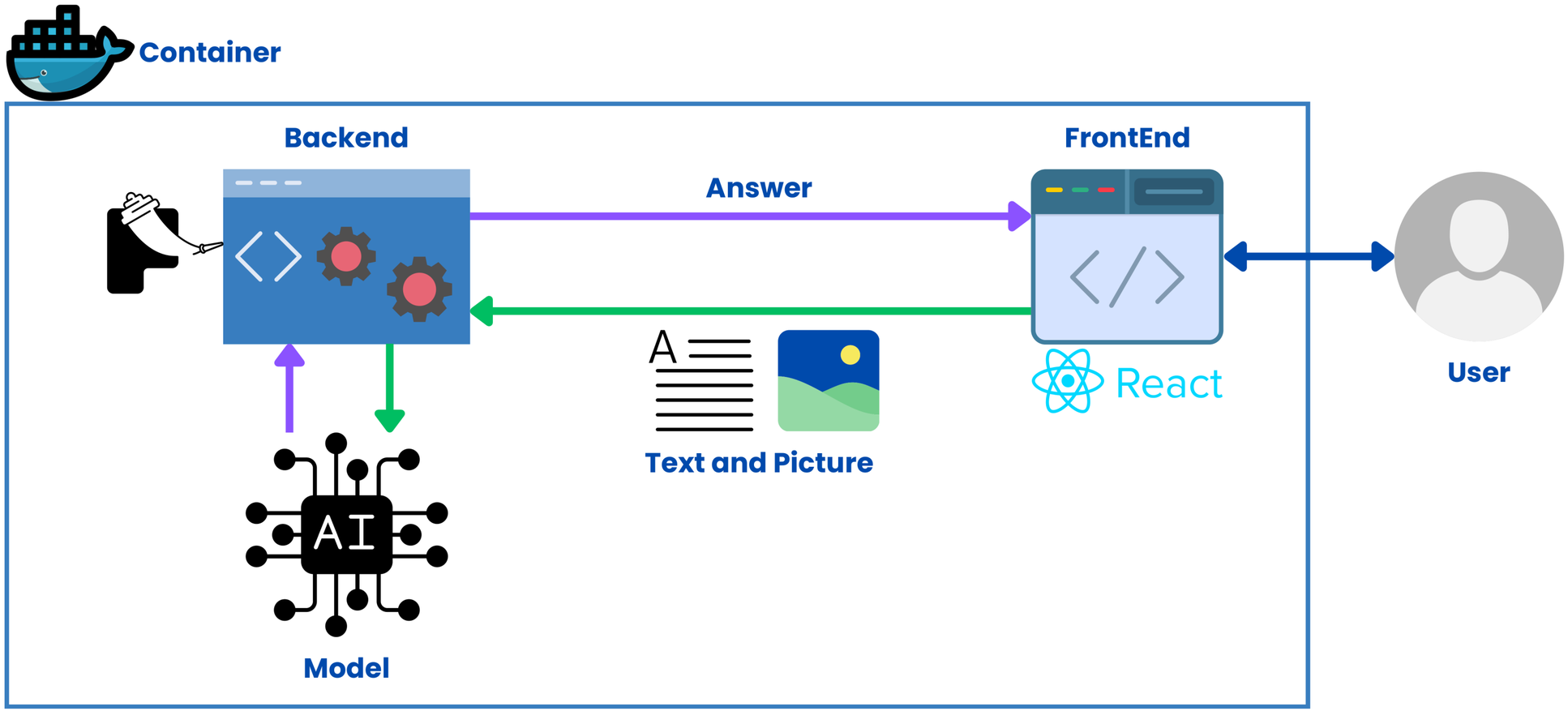
- Matcha: from huggingface
- Vicuna: from lmsys



CONDUCTED EXPERIMENTS



ARCHITECTURE



TASK DISTRIBUTION

Name	Main Task
Sitthiwat	Dockerization/UX
Parun	Models
Munthitra	Slide/Report
Pirunrat	Back End/Front End
Nathas	Models

THANK YOU

Question and Answer section