SkySearch: Find Anything

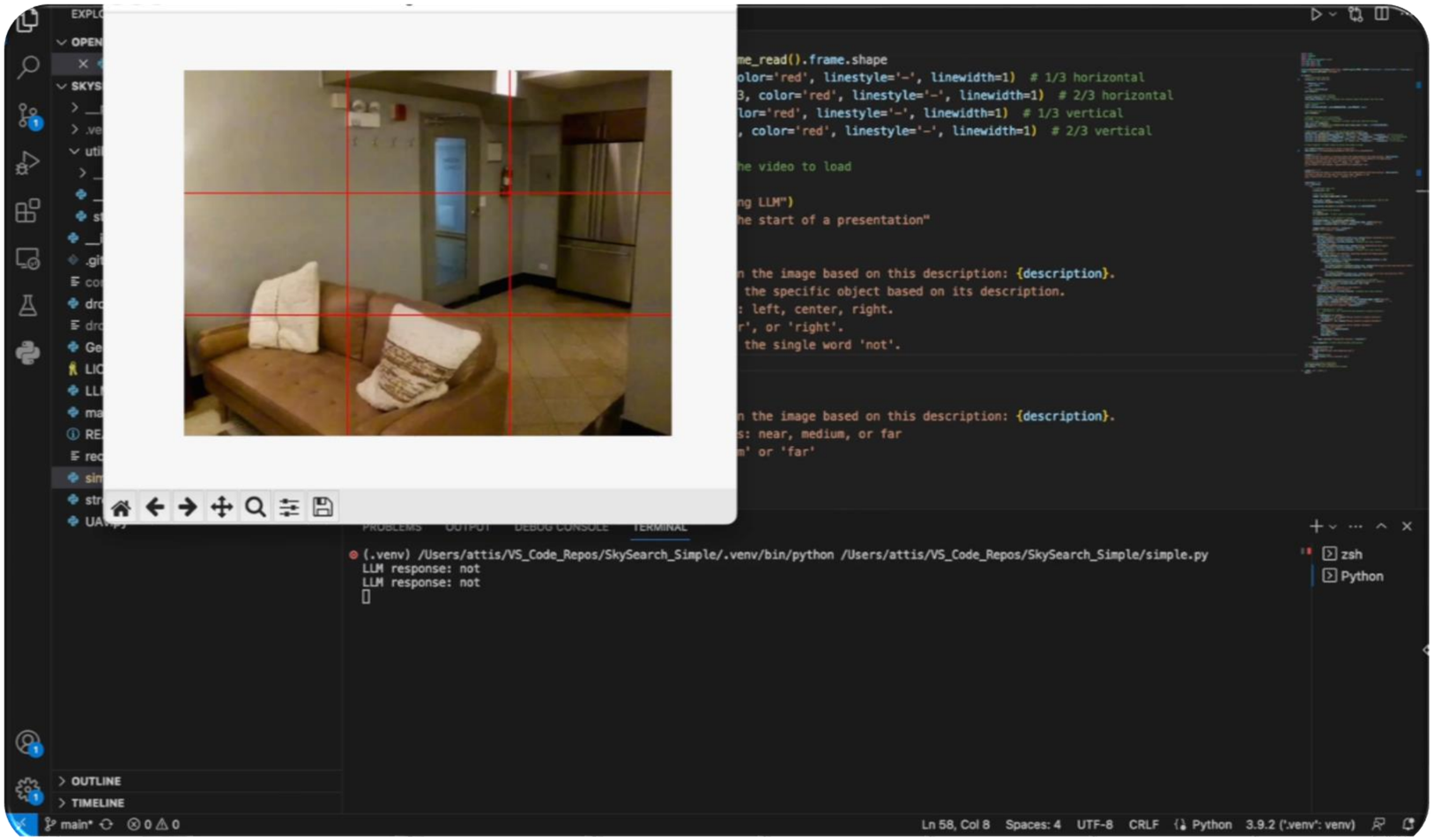
Generalized Object Search Using UAVs and Vision MLLMs

Zach Farahany, Joon Park, Duncan Calvert, Mohammad Ayan Raheel
University of Chicago M.S. in Applied Data Science
Robotics Capstone, December 2024

Let’s Get Started!

Prompt:

“Find an encouraging message to kick off the start of a presentation”



Agenda

1. Team Introductions
2. Research Problem & Opportunity
3. Robotics History
4. Project Methodology
5. Data Sets
6. Models and Architecture
7. Test Results
8. Conclusion & Future Research

Meet the Team

Duncan Calvert

Joon Park

Zach Farahany

Ayan Raheel

Problem Statement & Business Opportunity

Goal #1

Build a first-of-its-kind unmanned aerial vehicle (UAV) search system that can utilize rich text descriptions and logical reasoning to conduct open-vocabulary object detection.

Goal #2

Baseline the current SOTA vision MLLM models on their zero-shot search abilities

Potential Use Cases

Search & Rescue

Find a 15-year-old white male, 5'5" tall, 120lbs., with brown hair and brown eyes wearing a green jacket and black pants

Post-Disaster Survey

Find and waypoint any fallen trees or debris on roadways following a hurricane

Construction Site Safety

Find and waypoint any potential hazards or OSHA violations and provide reasoning for why they are an issue

Robotics: An Abridged History

1737

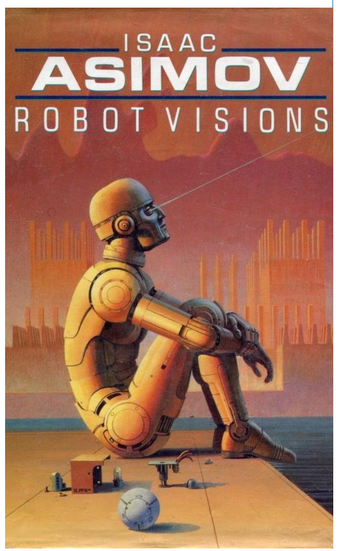
Jacques de Vaucanson's invents **The Flute Player**, an early humanoid clockwork automaton

1770

The Mechanical Turk was a fraudulent "automaton" that could play an opponent at chess and was wildly popular in Europe

1952

Isaac Asimov invents his **3 Laws of Robotics** and captivates the imagination of generations of engineers and inventors



1964

The Stanford Research Institute creates **Shakey,** the first mobile, general-purpose robot

2002

iRobot launches the **Roomba** robotic vacuum cleaner, the first everyday robot in homes

2014

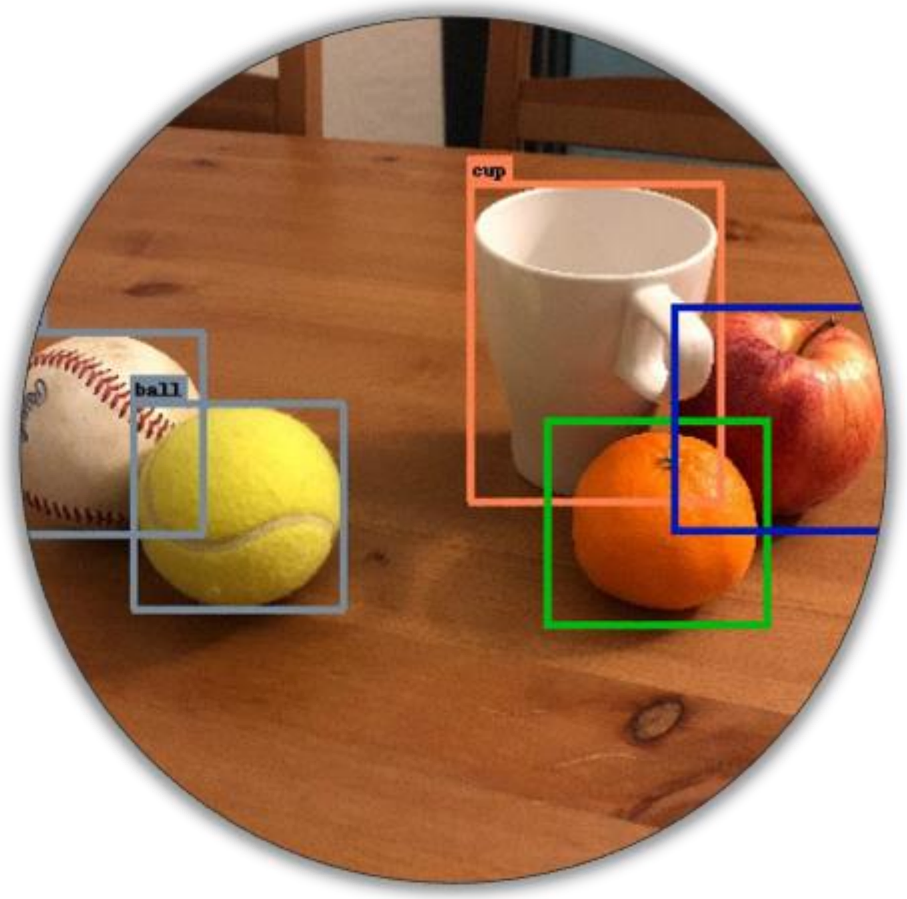
Zipline International Inc. is founded, the current largest international drone company

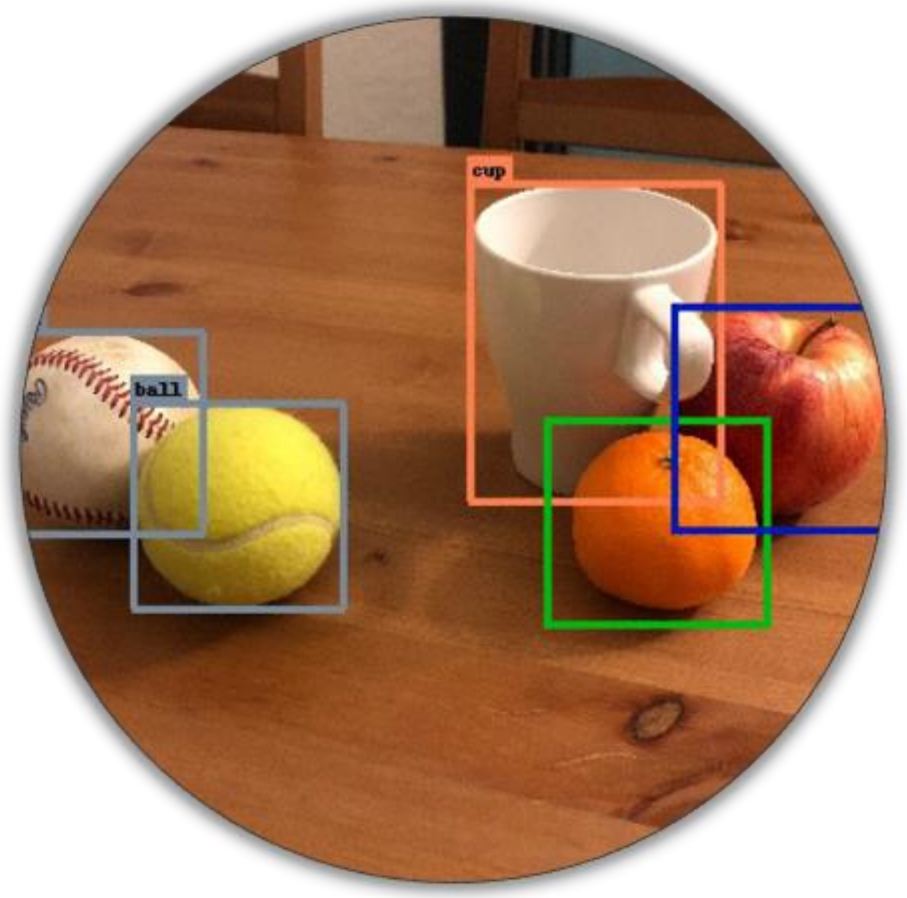
2023

Boston Dynamics integrates **Spot** with Open AI's GPT-4 model to allow their robot to speak and contextually understand its environment.

Object Detection Approaches

	Classic Models	YOLO-World	SkySearch
Models	R-CNN[1], YOLO[2], Feature Pyramid Networks	YOLO-World (CLIP + YOLO)	Vision MLLMs (Gemini 1.5, GPT-4o, Claude 3)
Object Vocabulary Size	Closed-Vocabulary (80)	Open-Vocabulary	Open-Vocabulary
Speed (V100)	~200 FPS	52 FPS	2-4 FPS
Task Scope	Object Detection	Object Detection	Image Captioning, Visual Question Answering, Image Reasoning, Object Recognition
Search Prompts	"Find the cup"	"Find the orange"	"Find the object that holds coffee"





1. Region-based convolutional neural network
2. You Only Look Once
3. Image source: https://apple.github.io/turicreate/docs/userguide/object_detection/

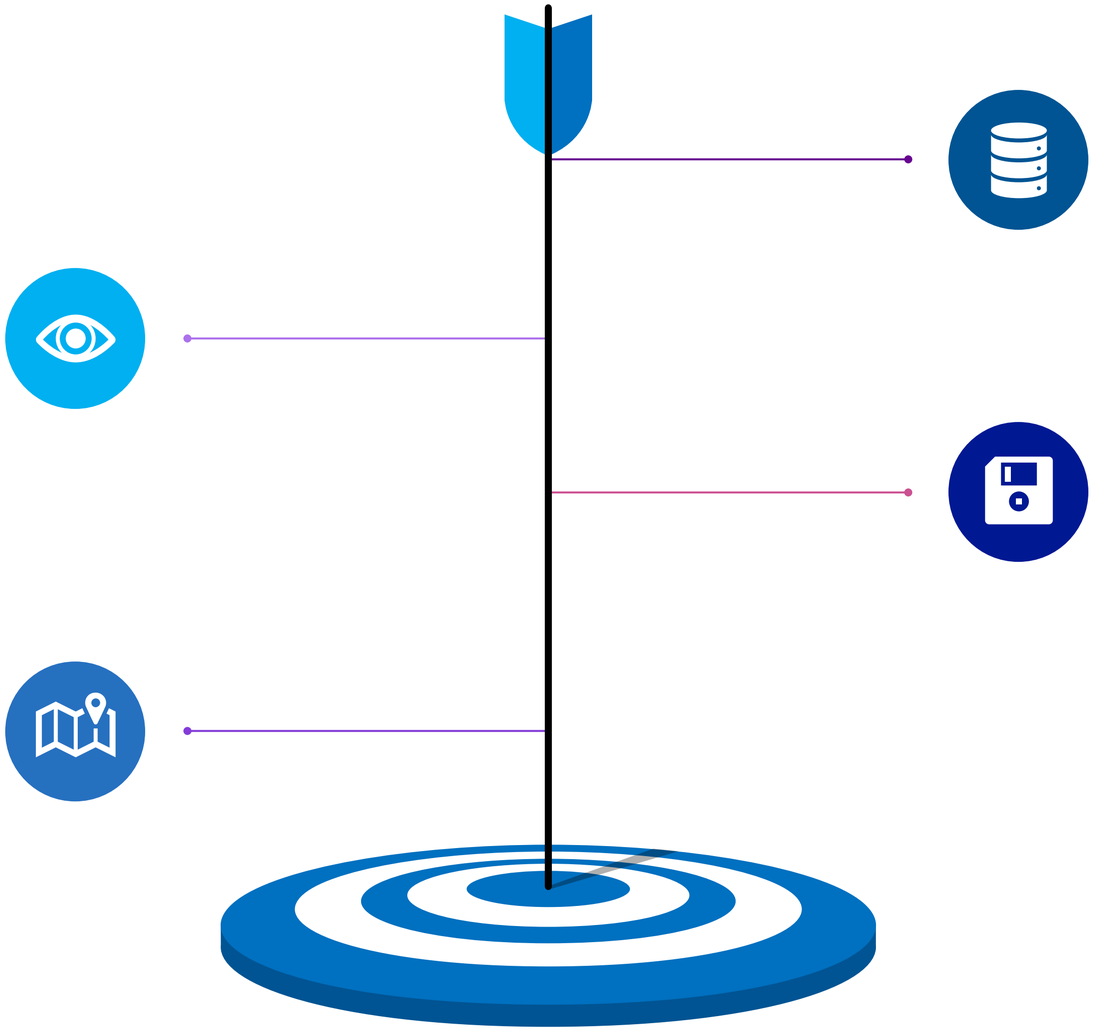
Project Components

Object Search System

A UAV Ground Control System that is able to search a set area using long-form descriptions and contextual logic

Waypoint Mapper

A clean application UI that displays the current drone video, controls, and a map of the location



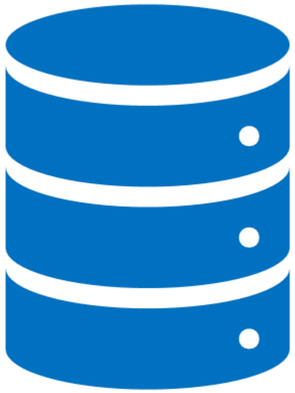
GLAD Data Set

A purpose-built evaluation data set for determining MLLM approximate object detection capabilities on open-vocabulary objects

Flight Audit

A flight log to allow for post-search auditing

Data Sets



Attributes	ImageNet	Common Objects in Context (COCO)	Grid Labeling Data Set (GLAD)
Publisher	Stanford Vision Lab	Microsoft	SkySearch
Image Count	14,000,000	330,000	300
Category Count	20,000	80	75
Object Context	Limited contextual information	Rich contextual information	Rich contextual information
Object Location	Bounding box	Bounding box	Image segment

Grid Labeling Approximation Dataset (GLAD)

Why is GLAD needed?

- Benchmark Data Contamination
- Approximate positioning and image segmentation vs. bounding box

Method

1. Segment image into 3x3x3 cube with horizontal, vertical, and depth
2. Label the following:
   - Name of the object
   - Is the object in the image (yes/no)
   - Horizontal position
   - Vertical position
   - Depth

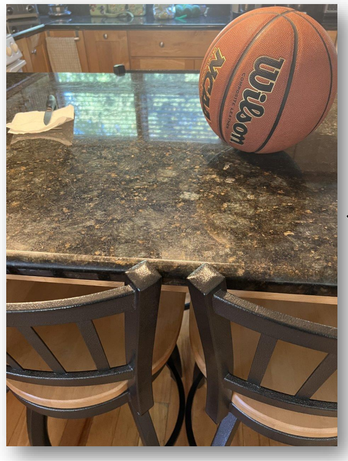
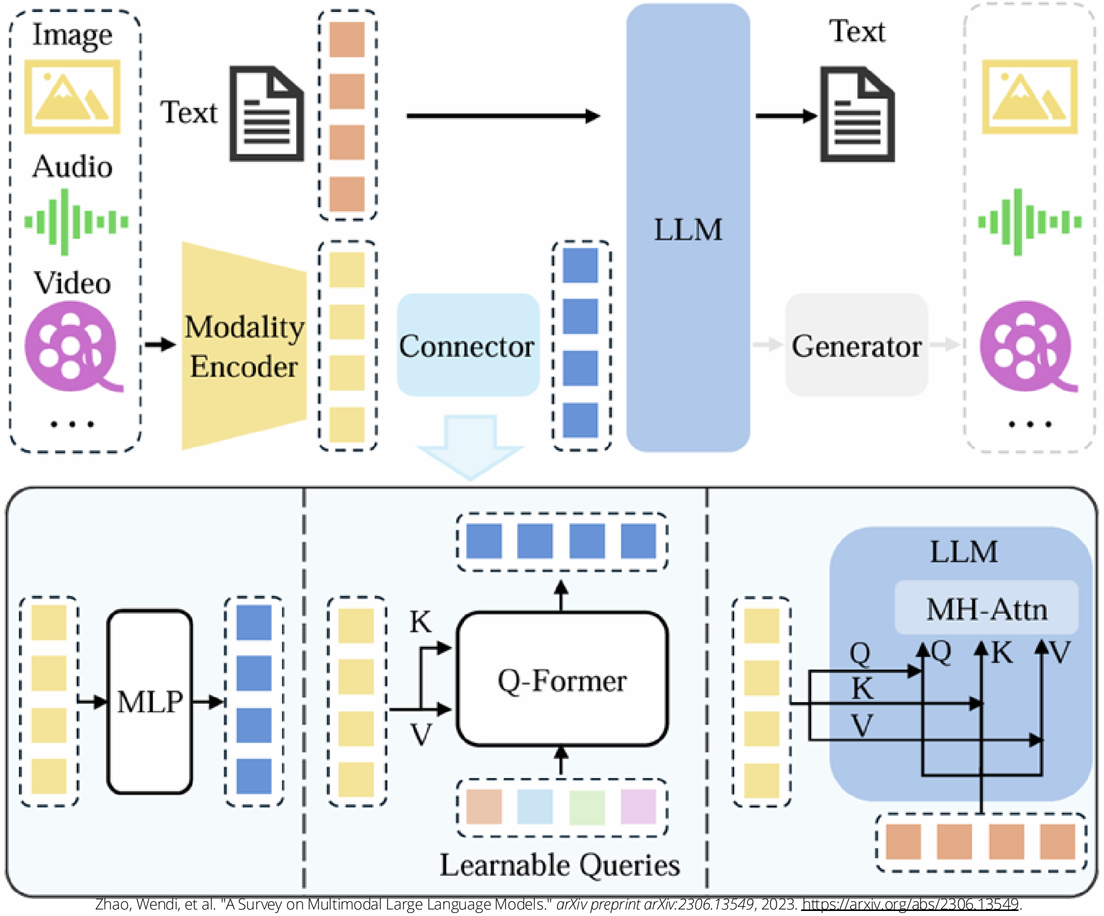


Image File Name	Item in Image	True Item Location	Annotator	Object in Frame
IMAGE_006.jpg	Apple	NaN	Zach	No
IMAGE_006.jpg	Globe	Left, Bottom, Medium	Zach	Yes
IMAGE_007.jpg	Basketball	Right, Top, Medium	Zach	Yes

Vision MLLM Architecture

Typical MLLM model architecture can be abstracted to three different modules:

1. **Pre-trained modality encoder**
   - Takes image as an input and converts it into embeddings that captures the important features and patterns in the image
   - Often accomplished using a vision transformer (ViT) or CNN based architecture.
2. **Modality Interface**
   - Combines visual and textual embeddings into a joint representation
   - The module is responsible for projecting the information into the space that LLM can understand efficiently
   - Often through a Q-former architecture
3. **Pre-trained LLM**
   - Interprets the fused representation created by the modality interface.
   - Generates natural language outputs based on combined vision and text information

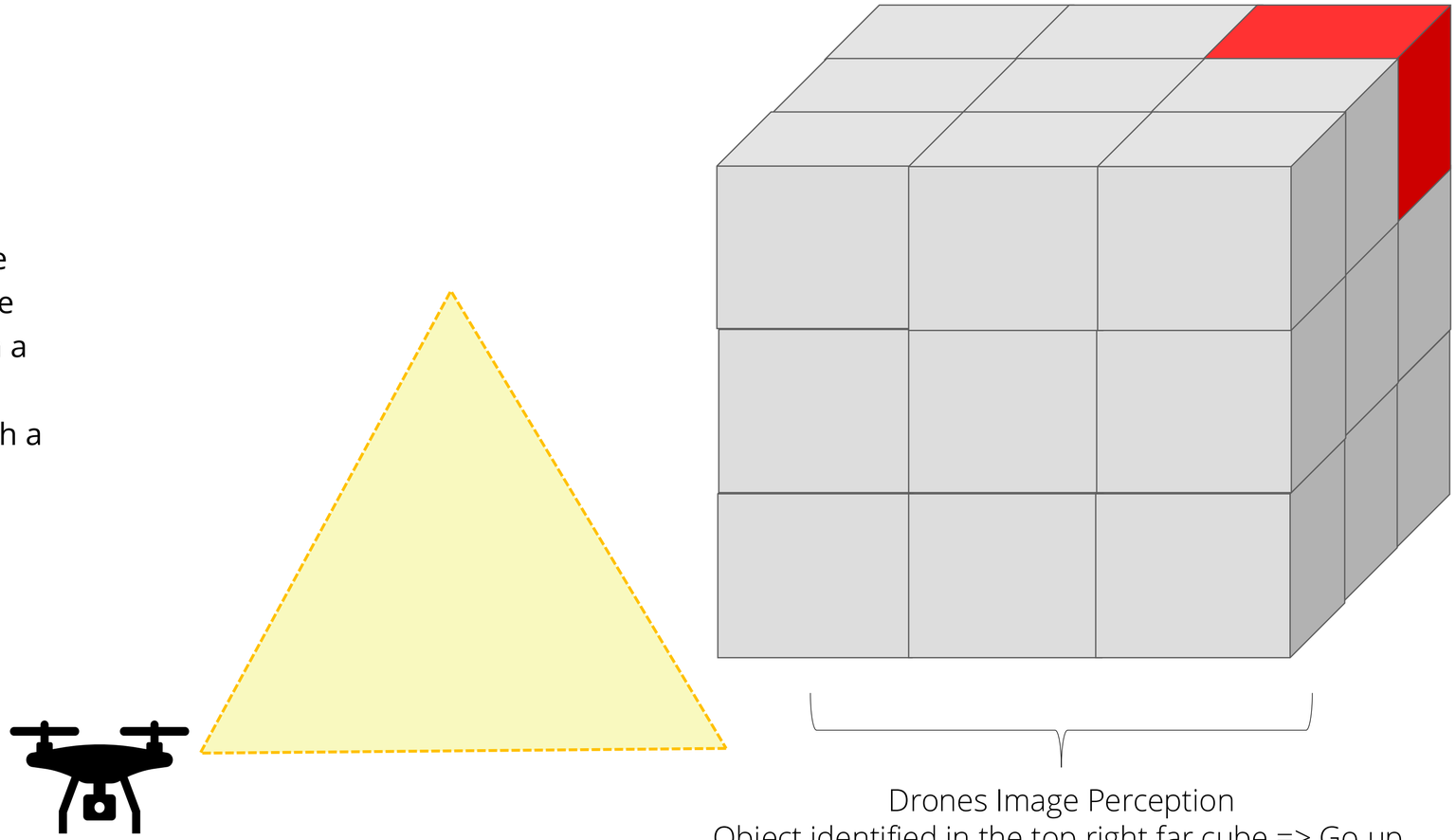


Zhao, Wendi, et al. "A Survey on Multimodal Large Language Models." *arXiv preprint arXiv:2306.13549*, 2023. https://arxiv.org/abs/2306.13549.

Approximate Object Detection

Translating an MLLMs image and prompt into actions

- In the prompt we ask the drone to split the drones image perception into a 3x3x3 cube
- Identify the target object's presence within a section with high accuracy
- Move toward the approximate location with a series of commands

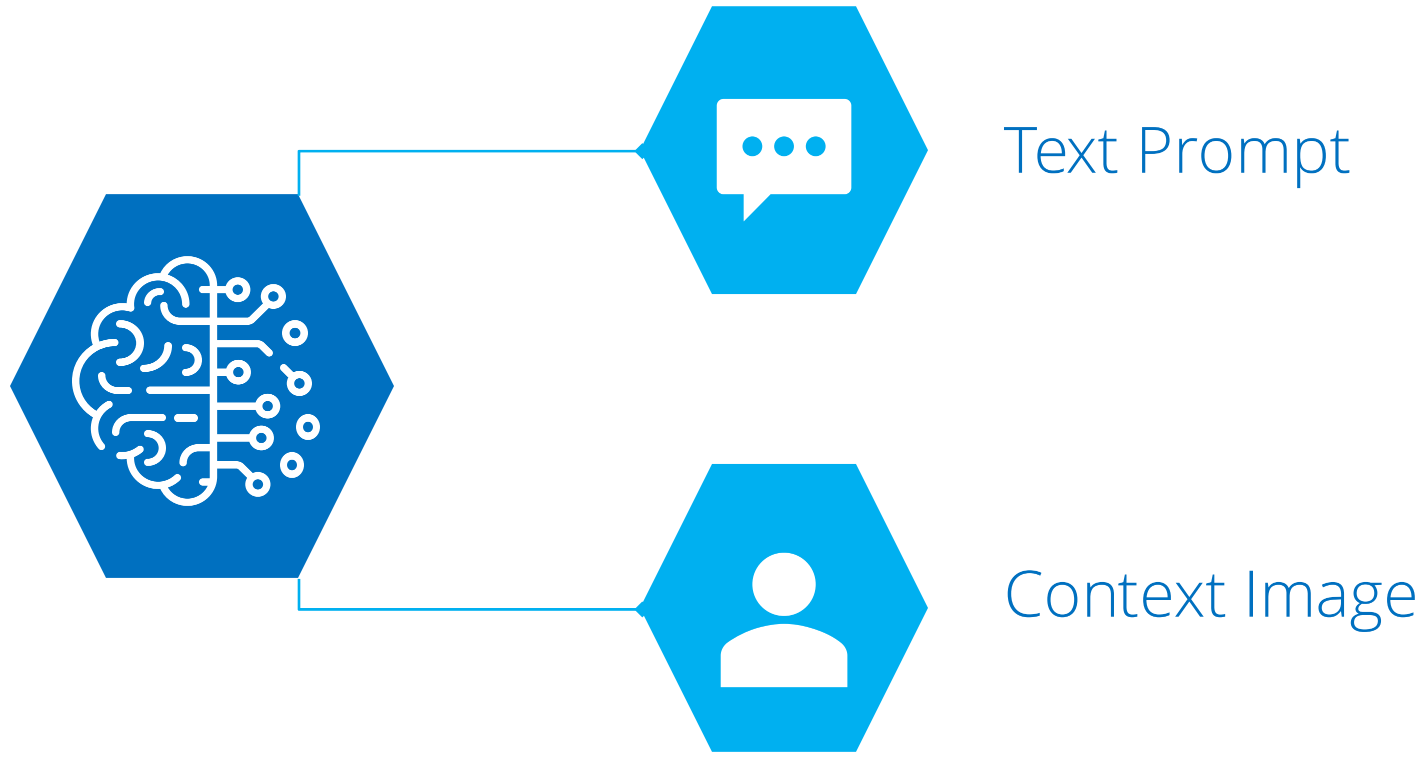


Drones Image Perception
Object identified in the top right far cube => Go up, right and forward

Zero-Shot Prompt Tuning

Why prompting tuning?

- LLM Object detection and recognition performance is highly sensitive to the prompt
- Tuning multiple prompts was necessary to achieve optimal results
- Identified 5 high-performing prompts and treated the choice of prompt as a hyperparameter




```
#########################################################################################################################
######################################################### PROMPT ########################################################

You are a vision model that is an expert at object detection. Locate the following object within the picture: {object}.

Instructions:
1. Determine if the object is in the image (yes) or if it is not in the image (no)
2. Divide the image into a 3x3 grid of equal size with the following combinations:
        • x-axis: left, center, right
        • y-axis: top, center, bottom
3. In addition to this categorize how far away the object is based on the following 3 categories
        • depth: near, medium, far
4. Based on the above, provide an exact 4 word response in this exact format {object in image, x-axis, y-axis, depth}
        • Example response: {yes, left, center, far}
        • Example response #2: {no, left, top, near}

Additional Instructions:
        • Exclude any additional words, punctuation, or capitalization.
```

The MLLM vision system prompt

Source: https://github.com/duncancalvert/SkySearch/blob/main/parameters/prompts.log

Evaluation & Testing - Overview

01

Open Vocabulary Object Recognition of MLLM Models

- Tested MLLMs (Gemini, GPT, and Claude) on the open-source ImageNet and COCO datasets

02

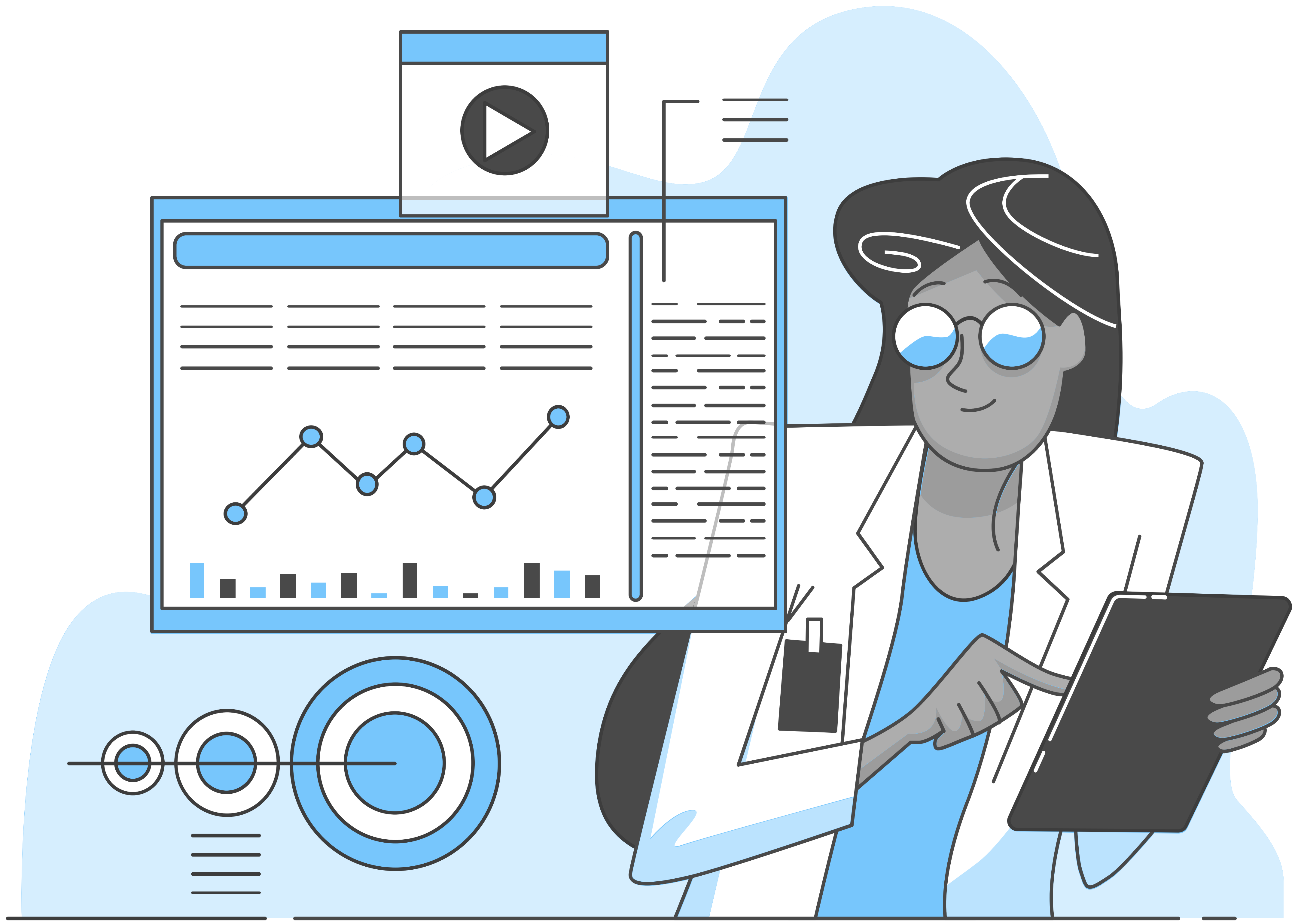
GLAD Approximate Location Testing

- Tested MLLMs on their ability to zero-shot object detection capabilities using approximate quadrant segmentation using the purpose-built GLAD data set

03

Live Flight Testing

- Ran a sequence of real-world tests in which a drone searched for an object using open vocabulary understanding



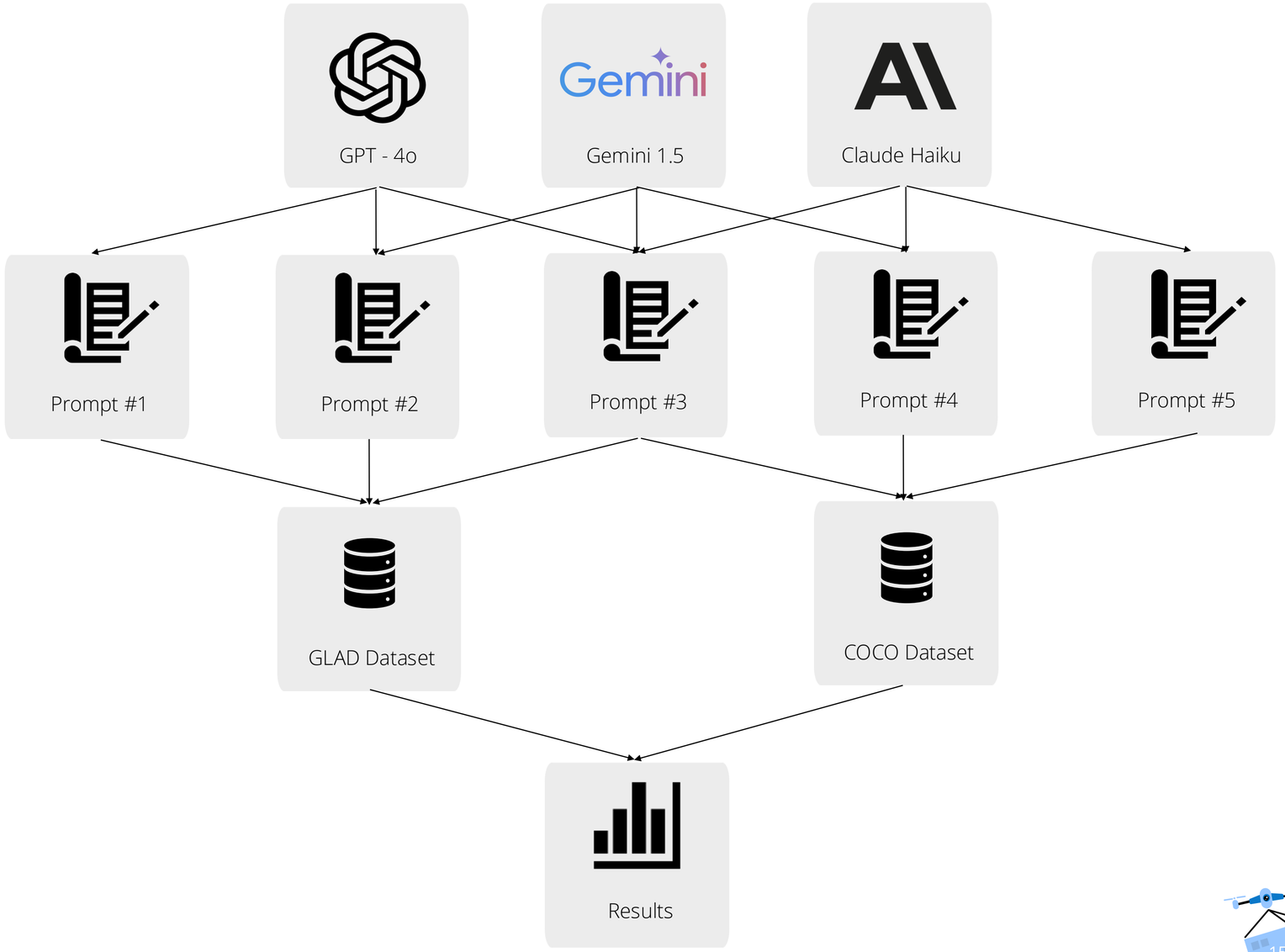
Evaluation Architecture

How do we evaluate MLLMs performance pre-flight?

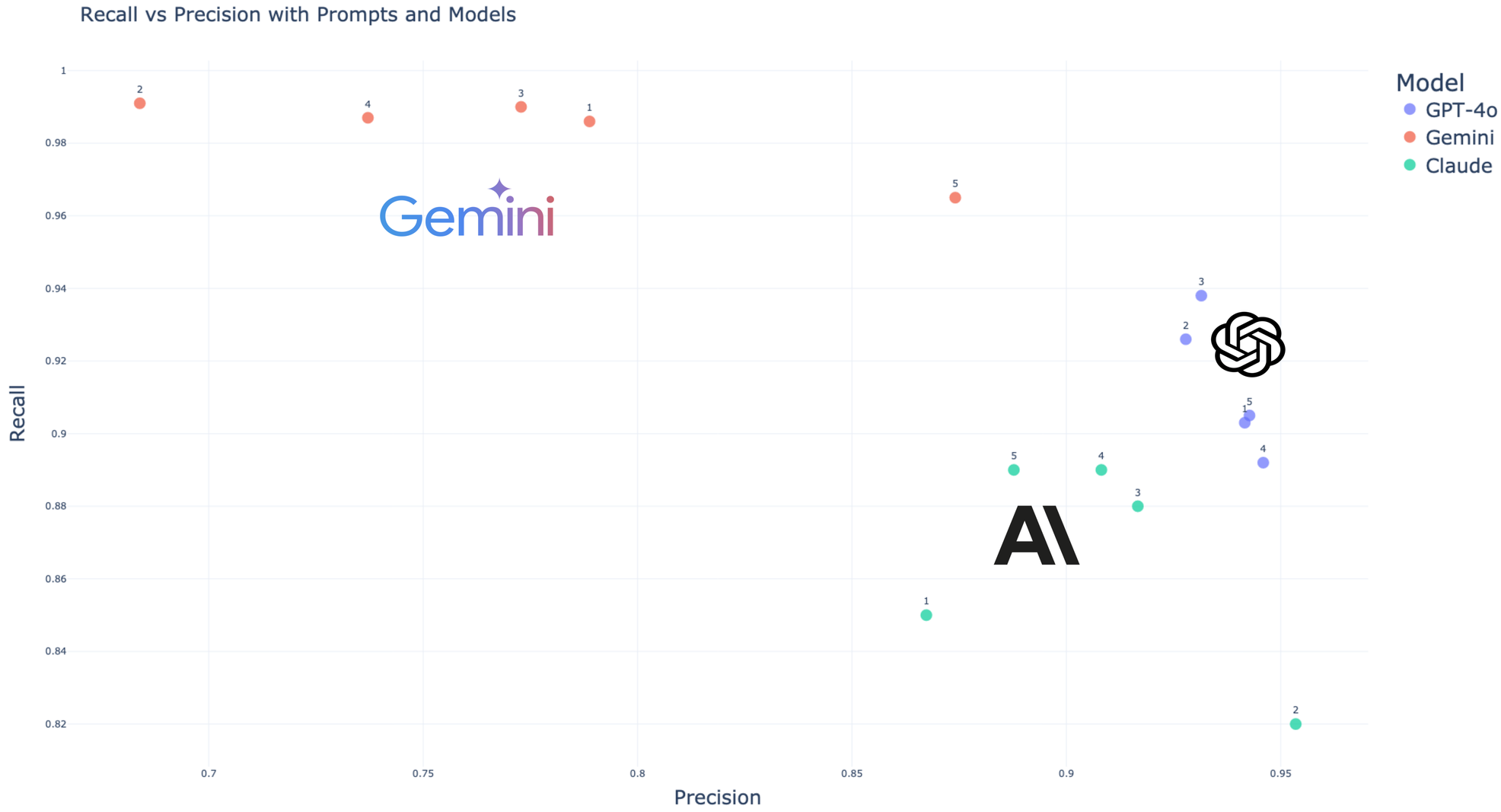
- Top multimodal LLMs evaluated across multiple optimized prompts
- Performance tested on two datasets: COCO and GLAD
- Metrics include accuracy and F1 score across all spatial dimensions

Key Takeaways

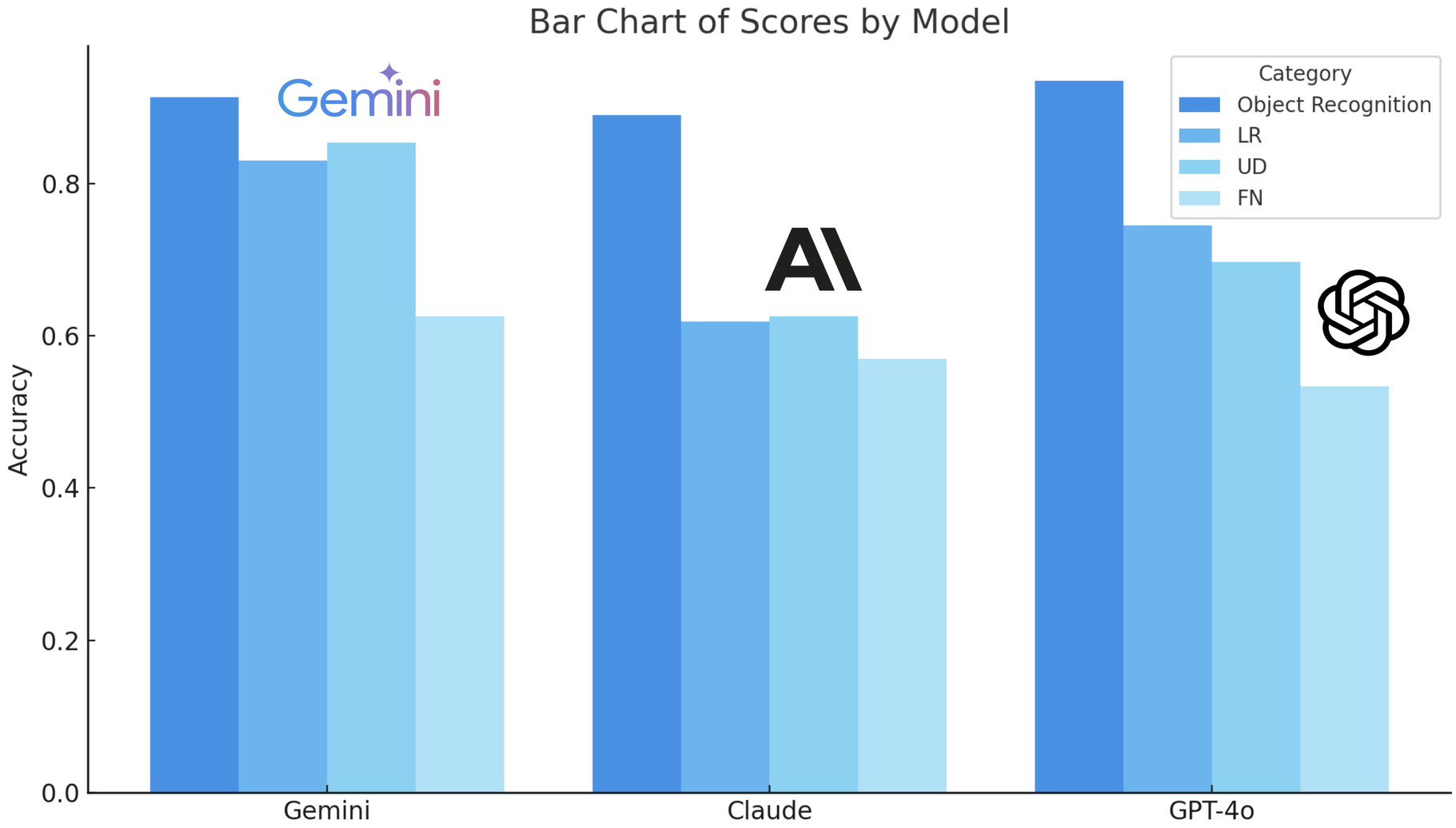
- This paradigm identifies the best LLM and prompt combination for object detection
- By testing on diverse datasets, we ensure robustness across real-world scenarios
- The results guide the selection of the most effective LLM/prompt for use in our drone



Evaluation & Testing – Generalized Object Recognition Results COCO



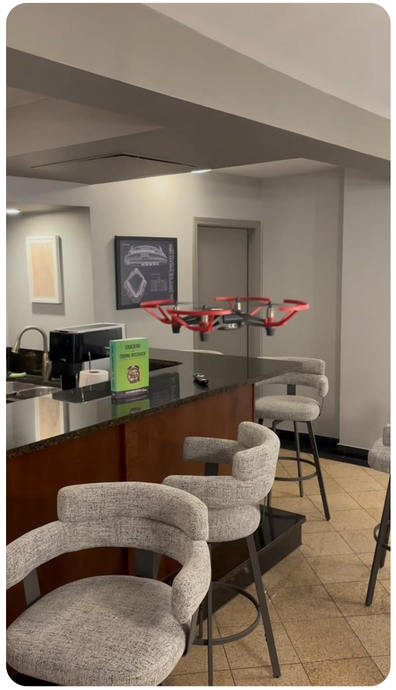
Evaluation & Testing – GLAD Approximate Location Object Detection Results

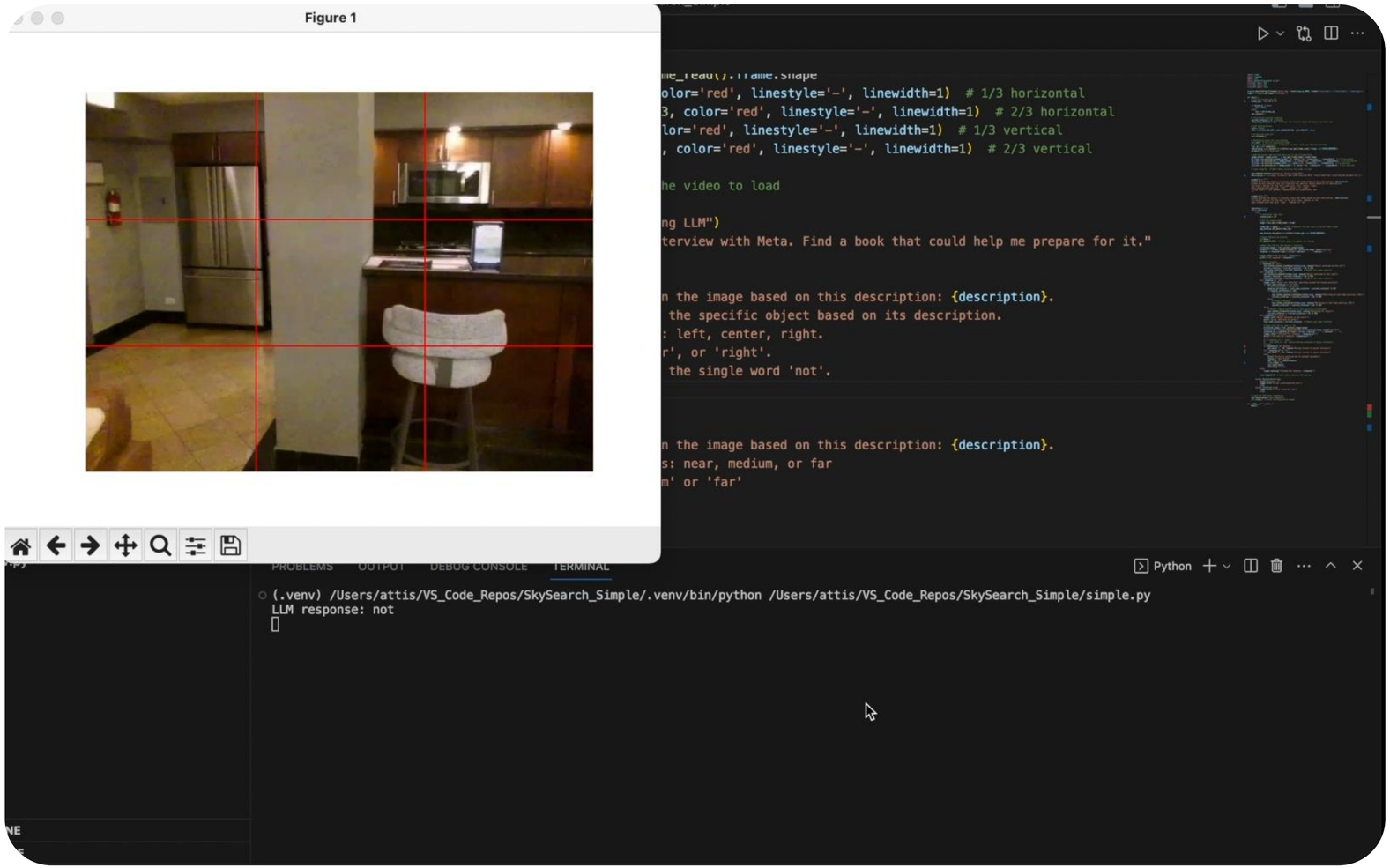


Evaluation & Testing – Flight Testing

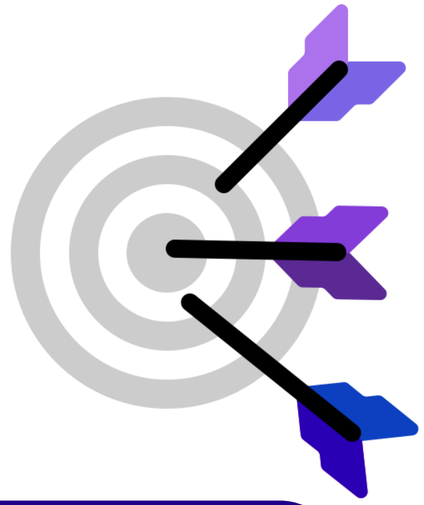
Prompt

”I'm about to have a tech interview with Meta. Find a book that could help me prepare for it.”





Conclusion & Findings



Speech-to-Text

Outcome

- Azure's Speech-to-Text advertised Word Error Rate (WER) is 5%, our testing showed closer to 15% which decreased overall search accuracy

Open-Vocabulary Object Recognition

Outcome

•The three MLLMS were able to perform object recognition with a 91.6% avg. success rate.

•The top performing model was GPT-4o:

- Accuracy: 93.5% across all classes.

GLAD Approximate Object Detection

Outcome

- The top performing MLLM was Gemini
  - Avg Accuracy: 77%
  - LR Accuracy: 83%
  - UD Accuracy: 85%
  - Depth Accuracy: 63%

Flight Testing

Outcome

- The flight system was successfully able to find non-COCO named objects using long form descriptions and logical context

Lessons Learned

Hardware

Use GPS-capable drones with longer flight distance (>100m), time (>13min), and gimbal control

GCP Model Garden

MLLM scoring pipeline throughput throttling limit (5 images/min)

Future Research

Optimized Swarm Search Routes

Optimized path planning based on a constrained geographic search space, swarm size, and real-time obstacle sensing

MLLM Vision Set Theory

- Experimental design to approximate object detection and recognition set size of MLLMs
- Experimental design to measure set size change over time for versioned MLLMs

UAV Swarm Polyocular Vision

Improved search accuracy through the use of a multi-drone perception system coupled with a voting layer

MLLM Vision Chain-of-Thought

Evaluation of MLLM vision accuracy on multi-step search processes that require reasoning

Resources

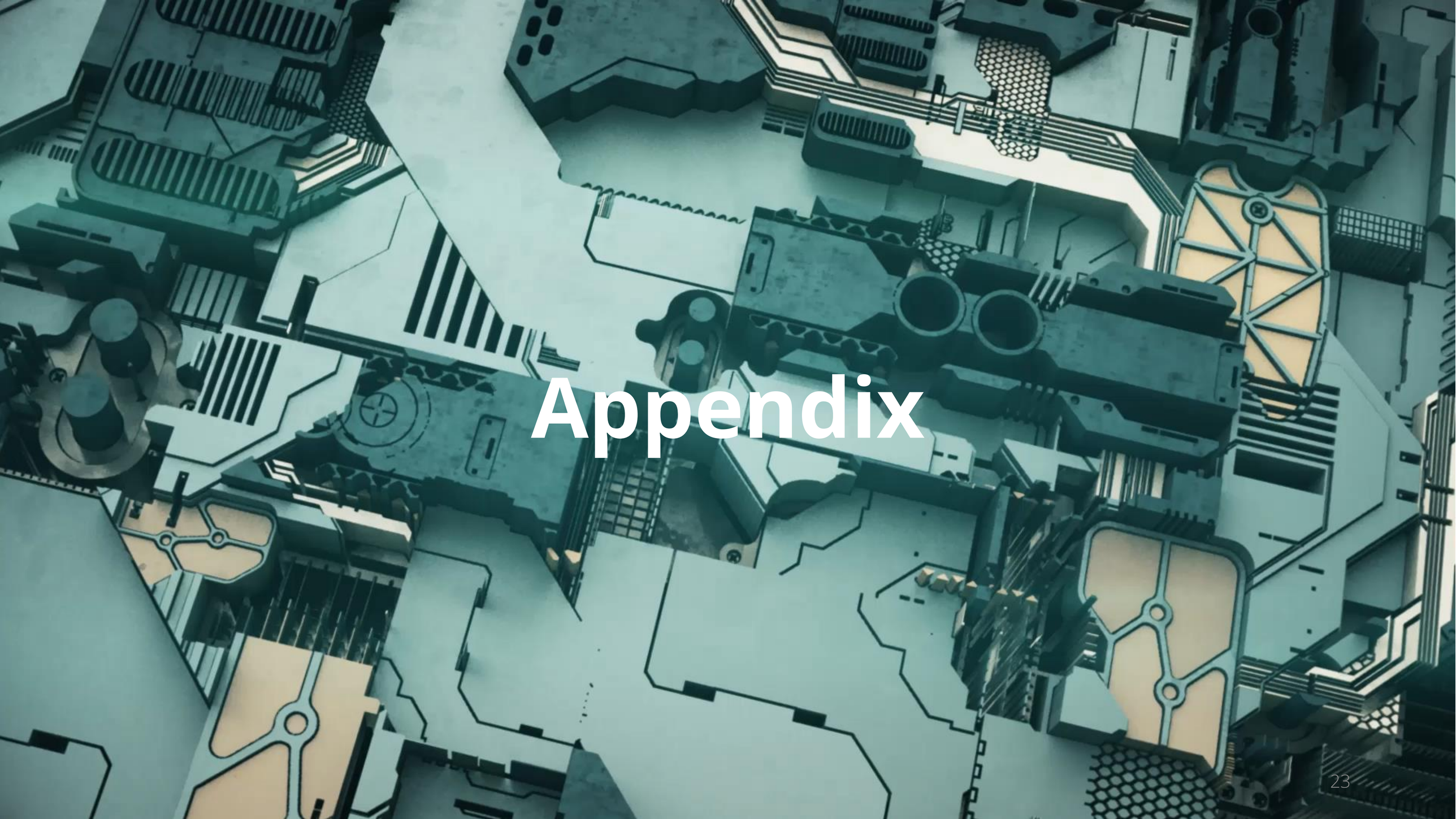
To learn more about SkySearch, please refer to the following resources:

Paper*

*Expected publication in January in arXiv

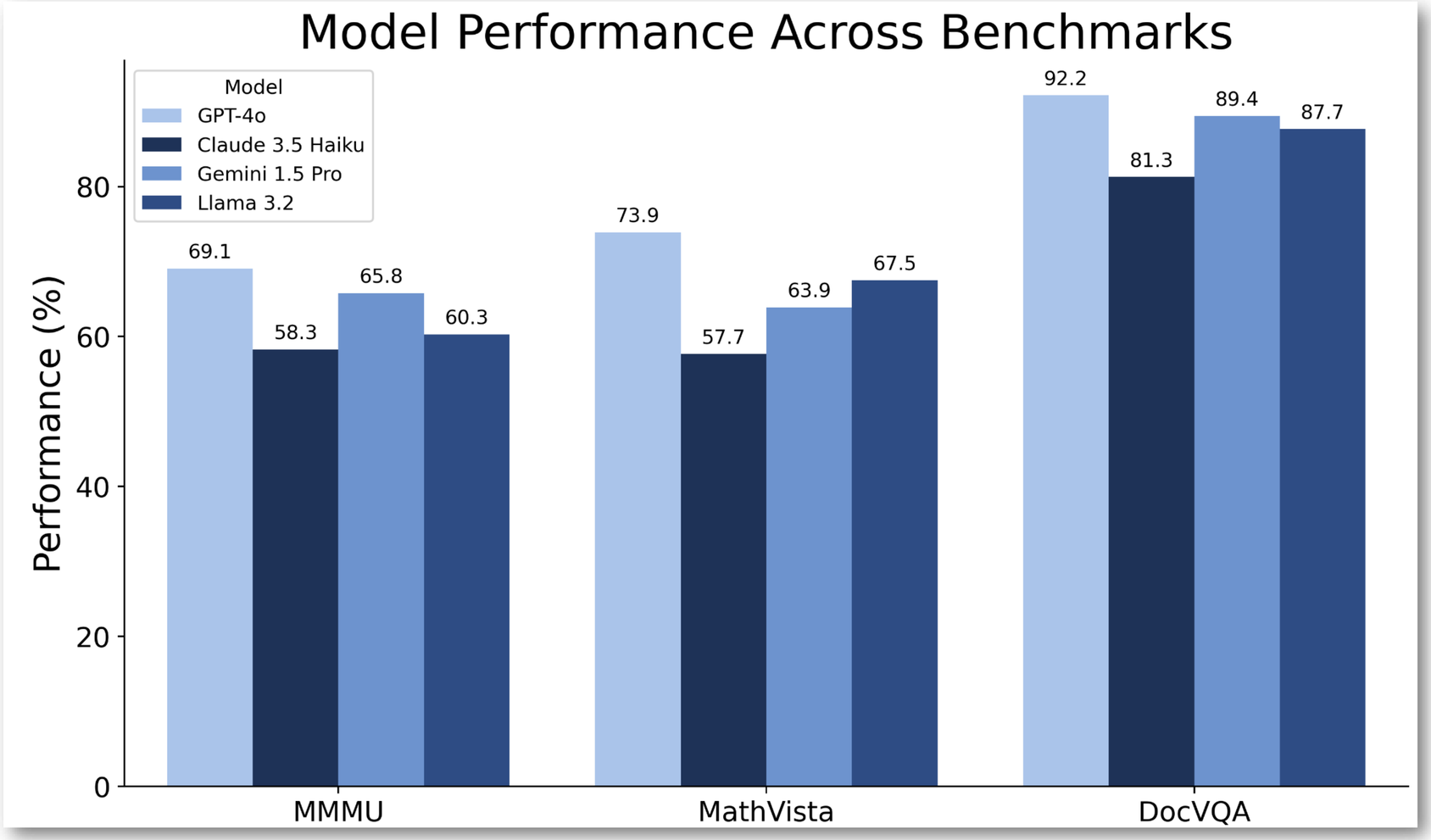
GitHub

https://github.com/duncancalvert/SkySearch



Appendix

Current Vision MLLM Leaderboard



https://mmmu-benchmark.github.io/

Current Vision MLLM Leaderboard

Benchmark	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro	Llama 3.2
Massive Multi-Discipline Multi-modal Understanding (MMMU)	69.1%	58.3%	65.8%	60.3%
MathVista	73.9%	57.7%	63.9%	67.5%
DocVQA	92.2%	81.3%	89.4%	87.7%

Literature Review

Object Goal Navigation

- Learning transferable visual models from natural language supervision, 2021
- ZSON: Zero-shot object-goal navigation using multimodal goal embeddings, 2023

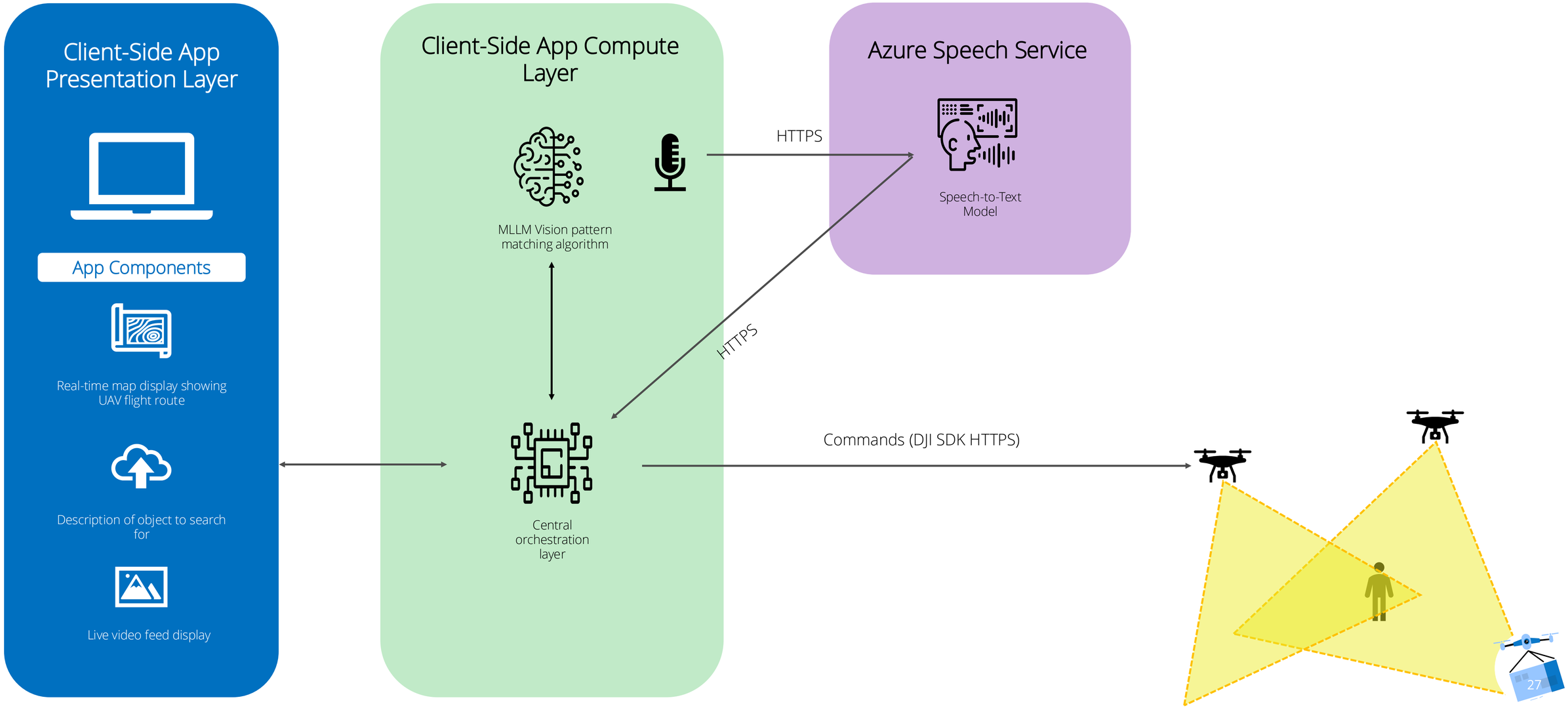
Image Data Sets

- Microsoft COCO: Common Objects in Context, 2015
- ImageNet Large Scale Visual Recognition Challenge, 2015
- Benchmark Data Contamination of Large Language Models: A Survey, 2024

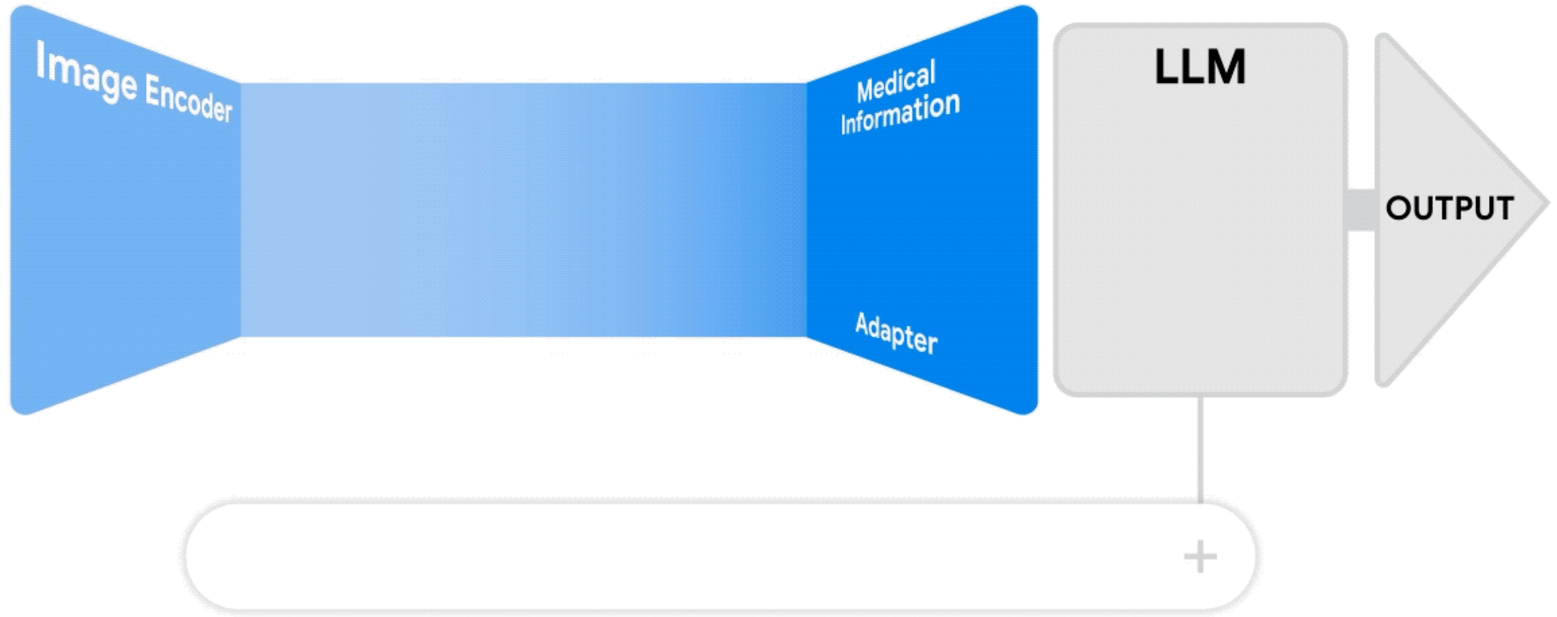
UAV Vision Modeling

- An Introduction to Convolutional Neural Networks, 2015
- You Only Look Once: Unified, Real-Time Object Detection, 2016
- Feature Pyramid Networks for Object Detection, 2017
- Mask R-CNN, 2018
- Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022
- Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks, 2022
- Semantic scene understanding with large language models on unmanned aerial vehicles, 2023
- Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023
- Semantic Scene Understanding with Large Language Models on Unmanned Aerial Vehicles, 2023
- YOLO-World: Real-Time Open-Vocabulary Object Detection, 2024
- Hello GPT-4o, 2024
- Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024
- Introducing the next generation of Claude, 2024

Live Flight Architecture



Vision MLLM Architecture



Multimodal Large Language Models (MLLMs) transforming Computer Vision | by The Tenyks Blogger | Medium

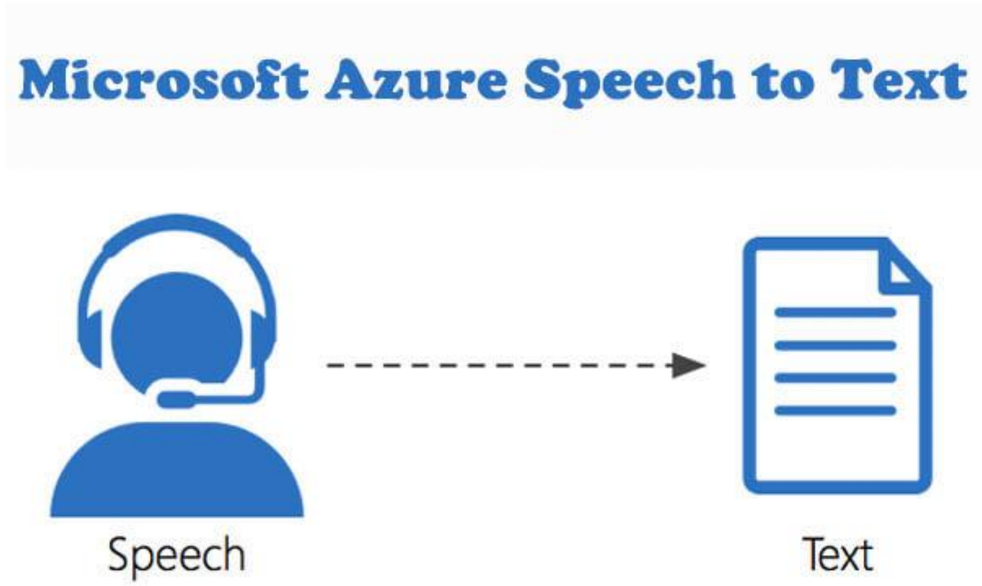
Speech-to-Text Models - Azure Speech Service

What is Azure Speech Service

- Azure Speech Service is a cloud-based speech recognition platform that enables developers to build speech-to-text capabilities into their applications.

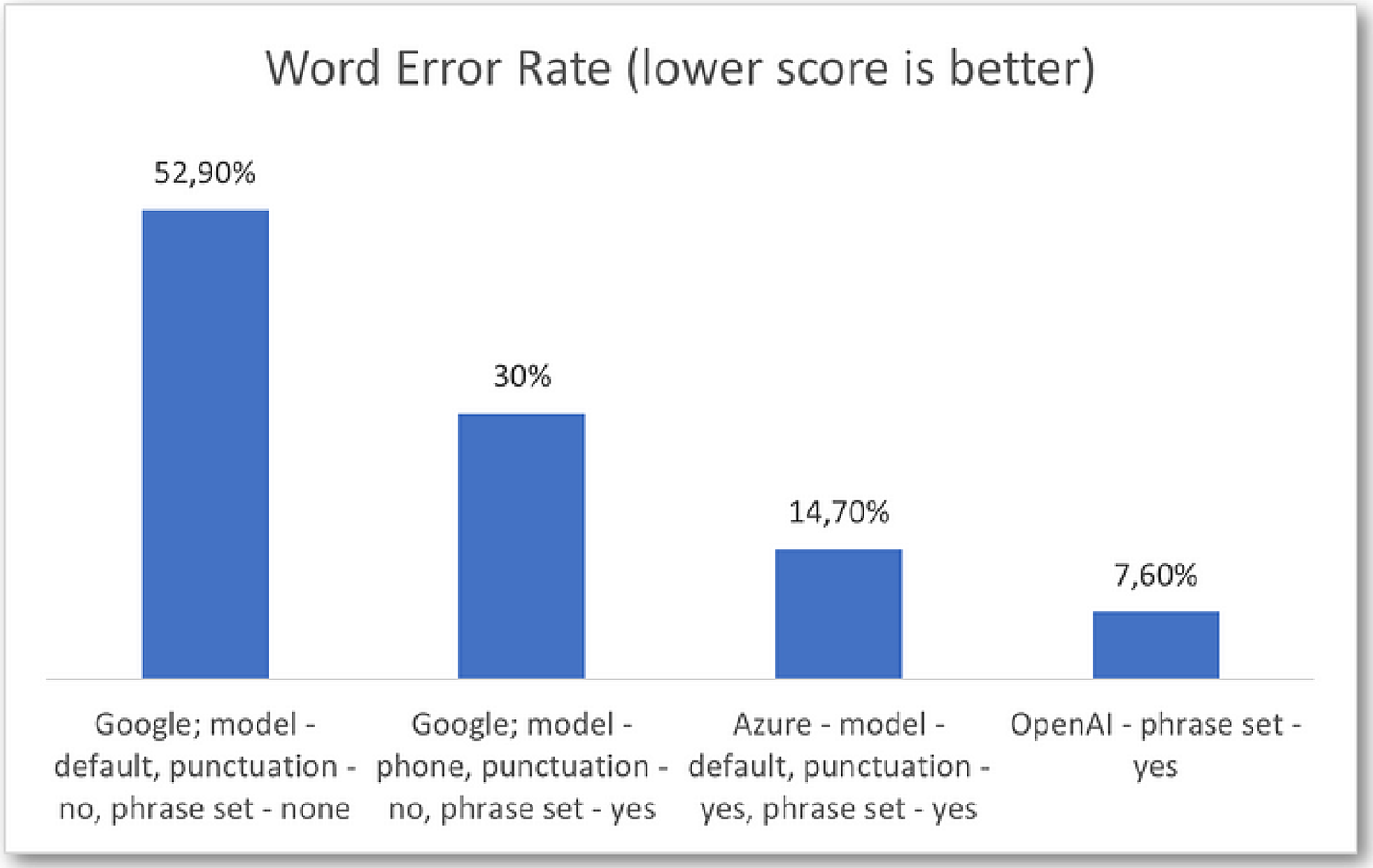
Key Features

- Converts audio to text in real-time or batch
- Supports 140 languages and dialects
- Integrates with other Azure services for a comprehensive speech solution



https://www.taygan.co/blog/2018/02/09/getting-started-with-speech-to-text

Speech-to-Text Models – Model Comparisons



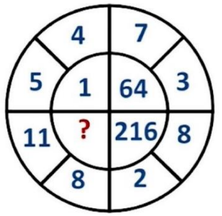
https://sstoitsev.medium.com/google-vs-azure-a-speech-to-text-battle-f740aa481e8e

Benchmark Dataset Examples

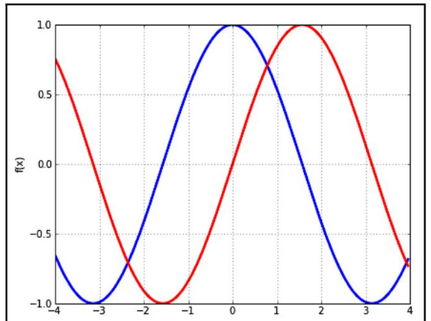
Massive Multi-Discipline Multi-modal Understanding

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> **(C) Minor seventh <*image 3*>** (D) Diminished sixth <*image 4*>	**Question:** ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? **Options:** (A) 0 (B) 0.2142 **(C) 0.3571** (D) 0.5	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** **(A)** $\int_0^{1.5}[f(x)-g(x)]dx$ (B) $\int_0^{1.5}[g(x)-f(x)]dx$ (C) $\int_0^{2}[f(x)-g(x)]dx$ (D) $\int_0^{2}[g(x)-x(x)]dx$
Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium	**Subject:** Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium	**Subject:** Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy
Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma **(C) Fat necrosis** (D) Silicone granuloma	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image* 1> **Option**: (A) Oppressor (B) Imperialist **(C) Savior** (D) Isolationist	**Question:** Find the VCE for the circuit shown in <*image 1*>. Neglect VBE **Answer: 3.75** **Explanation**: …IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

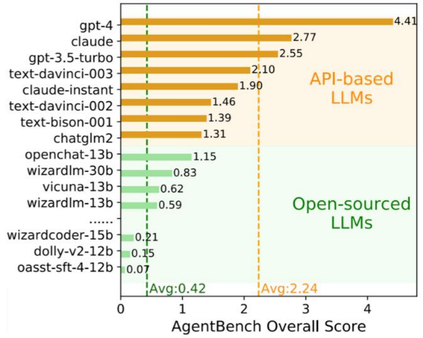
MathVista



Question: Find the missing value in this math puzzle.
Solution:
$(5 - 4)^3 = 1$
$(7 - 3)^3 = 64$
$(8 - 2)^3 = 216$
Similarly, $(11 - 8)^3 = 27$.
So the missing value is 27.
Answer: 27



Question: Which function is monotonic in range [0, pi]?
Choices:
(A) the red one (B) the blue one
(C) both (D) none of them
Answer: (B) the blue one



Question: What is the performance gap in the AgentBench Overall Score between the worst API-based LLM and the best open-sourced LLM?
Answer: 0.16

You Only Look Once (YOLO)

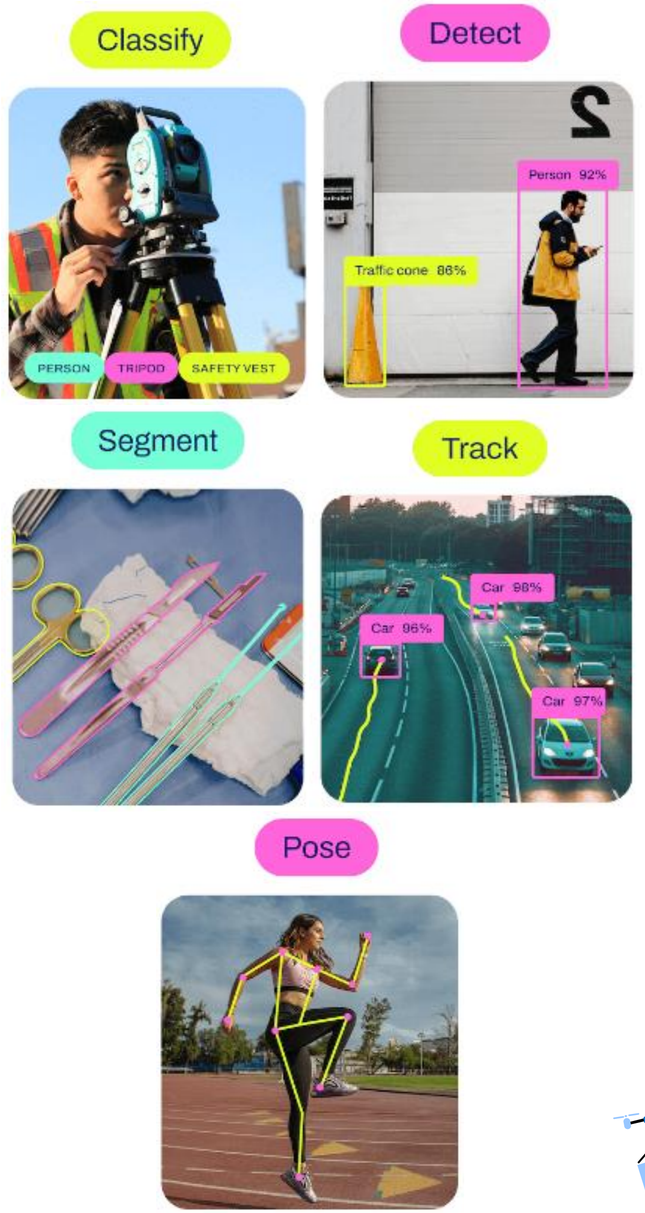
What is YOLO?

- YOLO is a real-time object detection system that detects objects in one pass through an image, hence the name "You Only Look Once". It's a popular computer vision (CV) package widely used for object detection tasks, particularly in applications that require speed and accuracy.

Key Features:

YOLO is known for its speed, accuracy, and ease of use, making it a go-to tool for developers, researchers, and businesses. Its architecture is designed to:

- Process images in real-time
- Detect objects in one pass
- Predict bounding boxes and class probabilities
- Handle object occlusion and context



YOLOv8 Ultralytics github

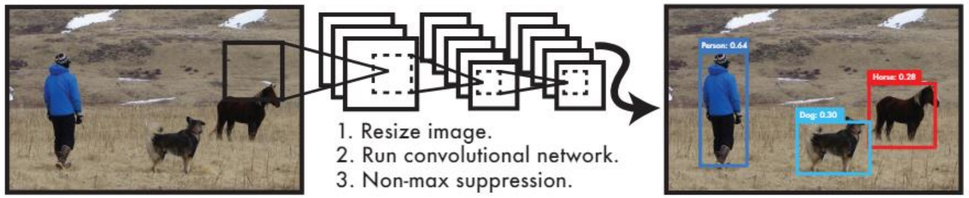
YOLO Detection Process

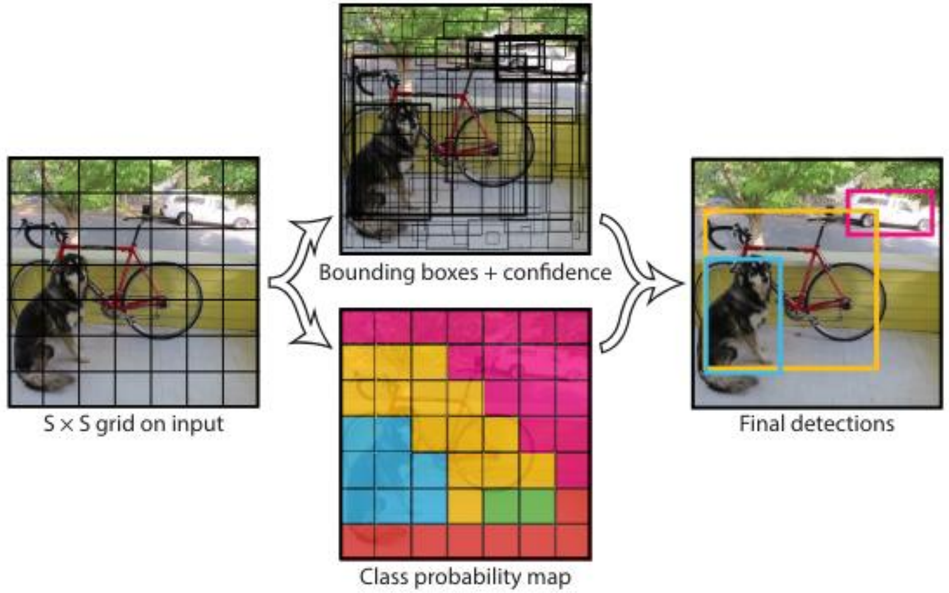
1. Image Resizing
- All input images are resized to a uniform dimension.
- This standardization helps the model to process varying image sizes and maintain consistency in detection scale.

2.Convolutional Network Processing(Next section)

3.Non-Max Suppression
- NMS reduces redundancy and filters out overlapping boxes by keeping only the box with the highest confidence score while discarding others that have a high overlap with it.

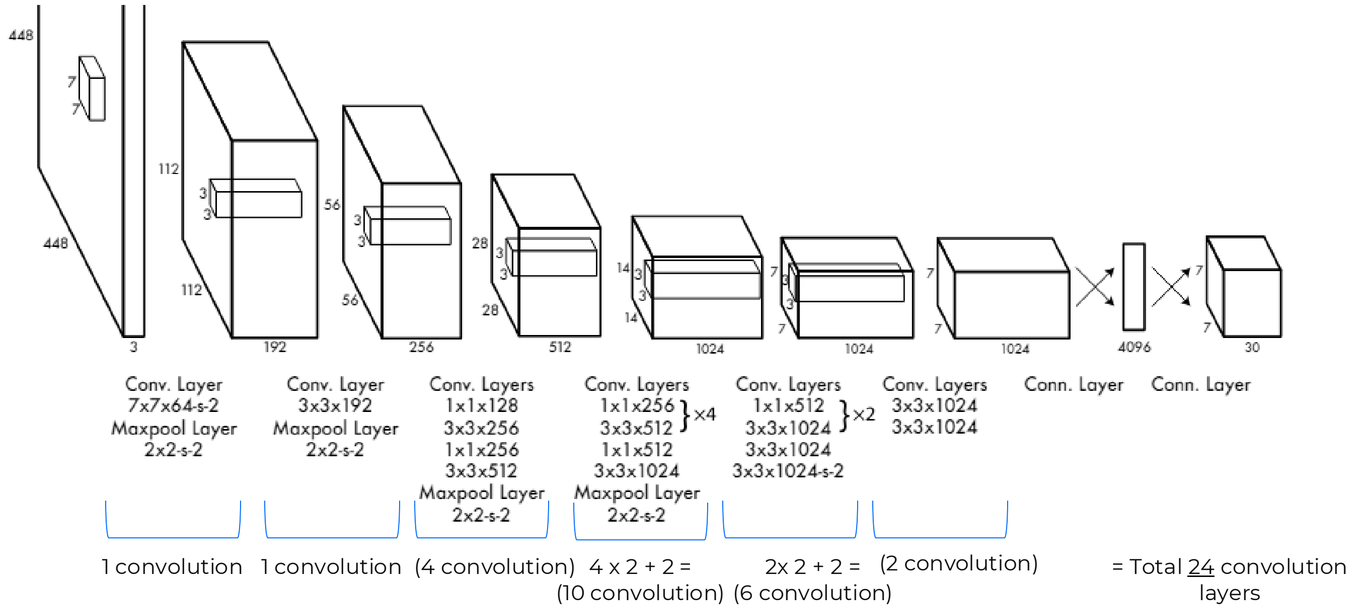
4.Confidence Thresholding
- Boxes with confidence scores below a certain threshold are discarded to improve the precision of the detection.





YOLO Detection System from the original paper

YOLO Architecture



YOLO Architecture from the original paper

- Resizes the input image into 448x448 (640x640 or 832x832 for YOLO v8)before going through the convolutional network.
- A 1x1 convolution is first applied to reduce the number of channels, which is then followed by a 3x3 convolution to generate a cuboidal output.
- The activation function under the hood is ReLU, except for the final layer, which uses a linear activation function.

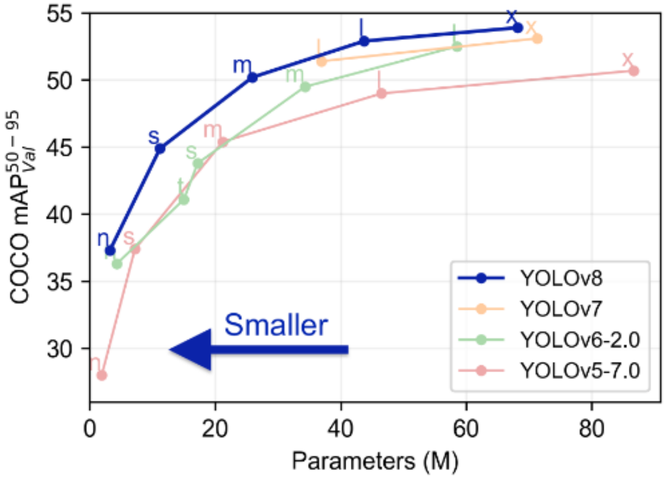
YOLOv8

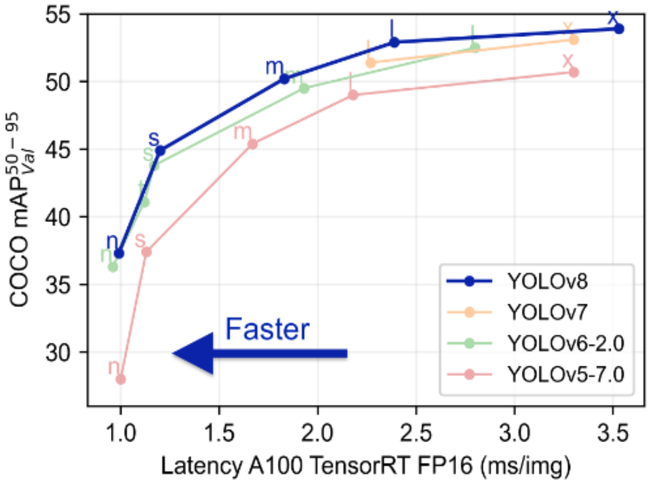
YOLOv8

YOLOv8 is the latest version of the YOLO family, released in 2022. It offers improved accuracy, faster inference, and new features like:

- Improved object detection architecture
- Enhanced anchor-free detection
- Support for instance segmentation and keypoint detection
- Improved performance on smaller objects and occluded objects

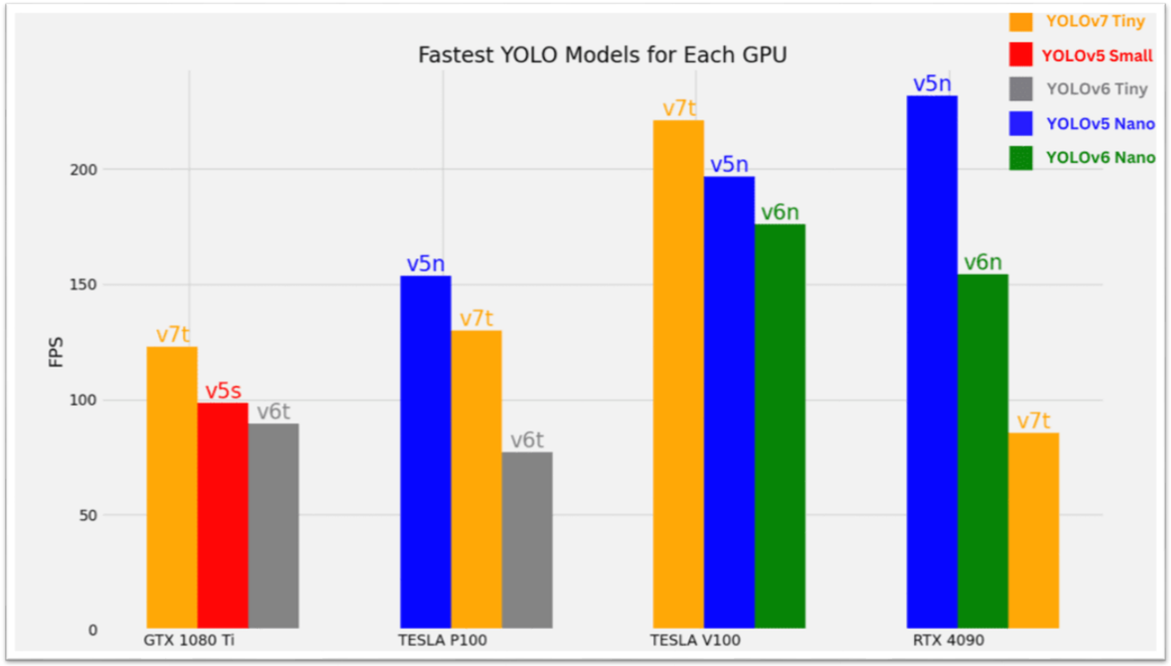
YOLOv8 is a significant upgrade, offering better performance and flexibility for various computer vision tasks.

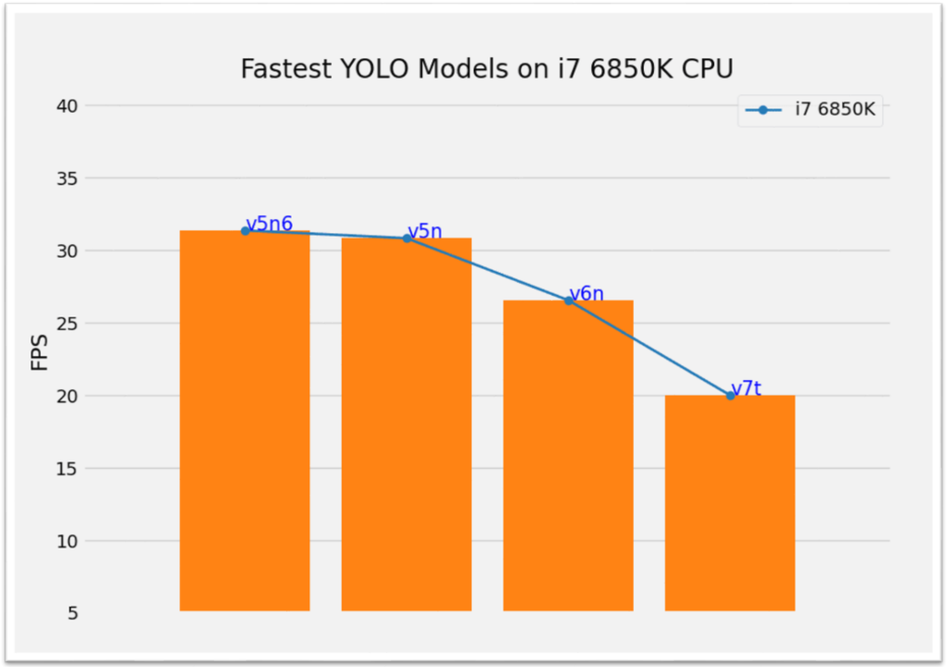




YOLOv8 Ultralytics github

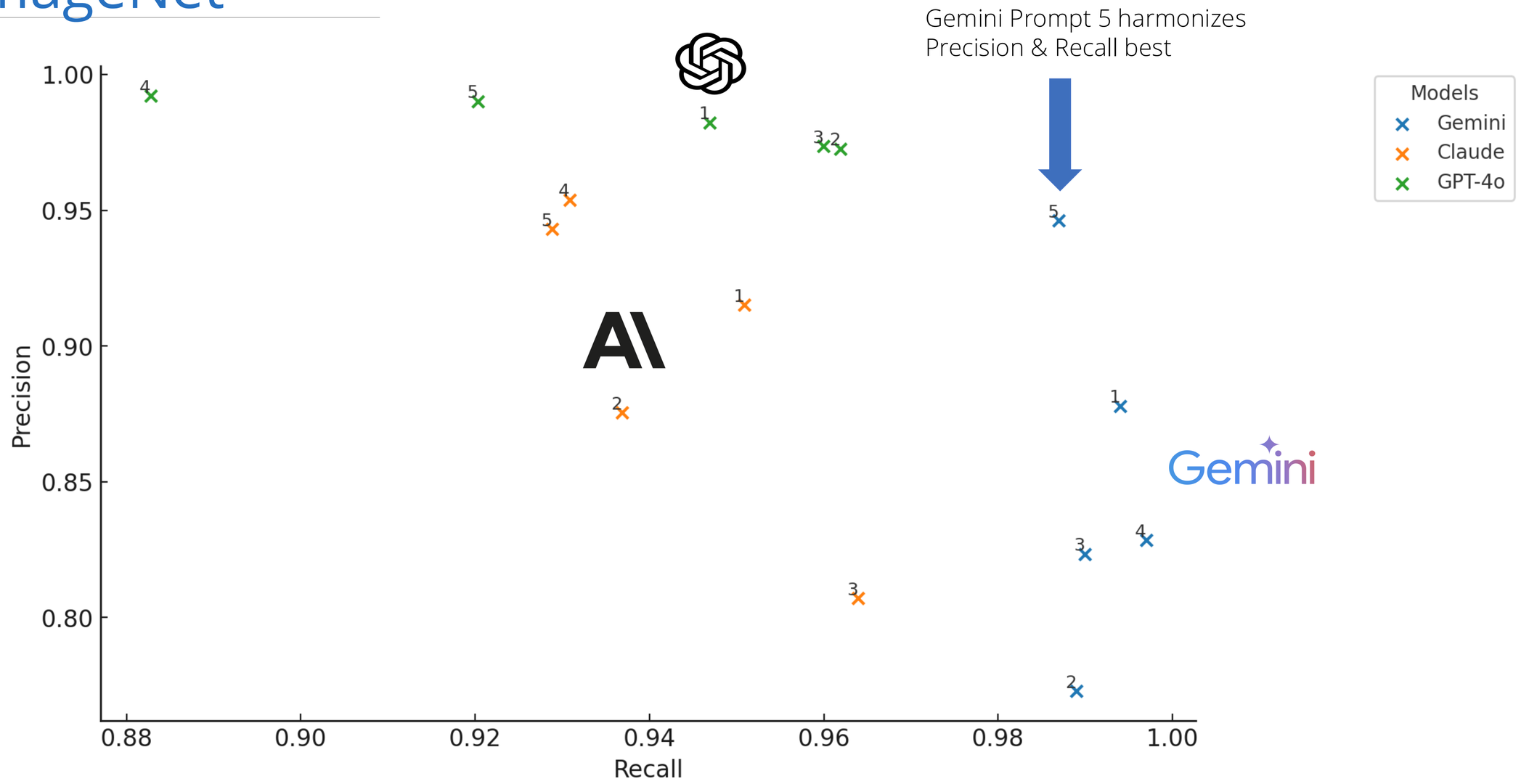
YOLO Performance (FPS)





https://learnopencv.com/performance-comparison-of-yolo-models/#Performance-Comparison-of-YOLO-Models-on-NVIDIA-Tesla-P100,-V100,-GTX-1080-Ti,-and-RTX-4090

Evaluation & Testing – Generalized Object Recognition Results ImageNet



UAV Search Ethics

“Technology is always a two-edged sword. It will bring in many benefits, but also many disasters.”

- Alan Moore

Benchmark Data Contamination

R-CNN

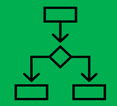
Panoptic Segmentation

UAV Flight Regulations & Compliance

What to Include

- Record the Date, Time and Duration of the flight from take-off till landing
- Location of the flight – the route the drone took in its flight covering all the major areas that saw the drone in the air
- Record the mission detail i.e., what was the purpose of the drone flight along with an inspection of the drone pre and post flight to ensure last line of defense

Other Considerations



- While the FAA does not require a flight log, it is highly recommended to create one for all the drone flights
- Adverse weather conditions can affect the flight of the drone and should be considered before any take off
- The choice to use either paper or an app to model the flight log is purely up to personal preference – each approach has its pros and cons

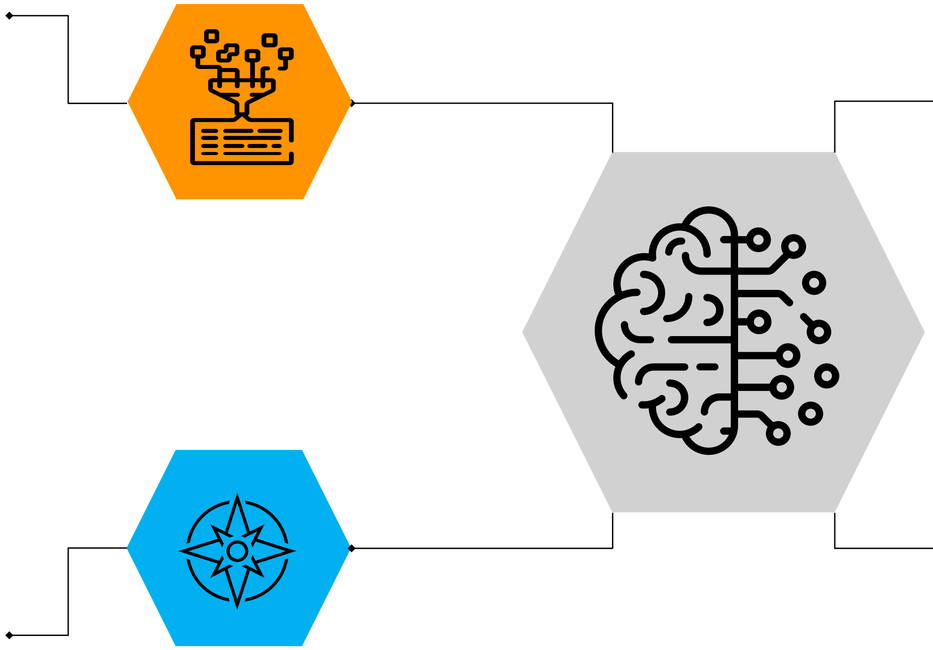
Models & Methodologies

MLLM Vision Models

- Gemini 1.5
- GPT-4o
- LLaMa 3.2

Speech-to-Text Drone Navigation

- GCP Speech-to-Text API
- Open AI Whisper



- Python Linear Programming (PuLP)
- Reinforcement Learning

Swarm Routing

- YOLO v11
- YOLO World

Non-MLLM Vision Comparison Models

Old Versions of Slides

Meet the Team – Duncan Calvert

Professional profile
- Lead Quantitative Analytics Specialist, Wells Fargo

Education
- M.S. Candidate in Applied Data Science, UChicago
- Post Bacc in Software Engineering, UC Berkeley
- B.A. in History and Political Science, UCSB

Personal interests
- Welding, blacksmithing, ceramics, industrial arts
- Photography

Primary and Secondary role within the Team
- Modeling Committee

Meet the Team – Joon Park

Professional profile
- Data Science Intern at Branch

Education
- M.S. Candidate in Applied Data Science, UChicago
- B.S. in Commerce, UVA

Personal interests
- Bouldering, Skiing, Reading, Gaming

Primary and Secondary role within the Team
- Coding and Navigation Committee

Meet the Team – Zach Farahany

Professional profile
- Data Scientist I at Badger Meter Inc.

Education
- M.S. Candidate in Applied Data Science, UChicago
- B.S. in Math and Computer Science, Marquette University

Personal interests
- Music, Programming, Reading, Language Learning

Primary and Secondary role within the Team
- Project Management Committee

Meet the Team – Mohammad Ayan Raheel

Professional profile
- Strategy Consultant at FSMS, Dubai

Education
- M.S. Candidate in Applied Data Science, UChicago
- B.S. in Economics & Political Science, LUMS

Personal interests
- Soccer, Formula 1, Blues & Jazz Music, Mixed Martial Arts

Primary and Secondary role within the Team
- Project Management Committee

Robotics: An Abridged History

1737

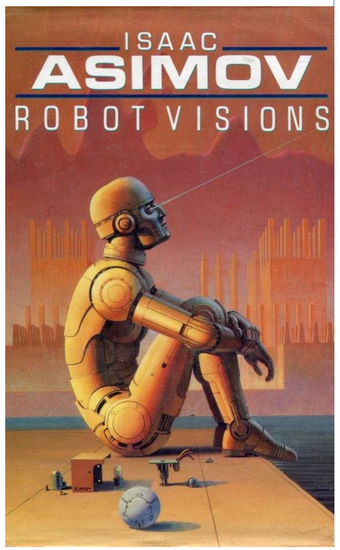
Jacques de Vaucanson's The Flute Player, an early humanoid clockwork automaton, the life-sized figure could play 12 songs via the use of bellows and pipes

1770

The Mechanical Turk was a "machine" that could play an opponent at chess and was wildly popular in Europe. The Turk was fraudulently advertised as an automaton but was in fact a human chess player

1952

Isaac Asimov invents his **3 Laws of Robotics** and captivates the imagination of generations of engineers and inventors with classics such as *I, Robot*.



1964

Shakey is created at the Stanford Research Institute. It is the first mobile, general-purpose robot that combines logical reasoning and physical actions

2002

iRobot launches the **Roomba** robotic vacuum cleaner. It would go on to sell 40 million units and become the first everyday robot in the home

2014

Zipline International Inc. is founded, the current largest international drone delivery company with 800,000 commercial deliveries operating in Africa, Asia, and North America

2023

Boston Dynamics integrates **Spot** with Open AI's GPT-4 model to allow their robot to speak and contextually understand its environment. This builds on their previous success with BigDog, Atlas, and WildCat

Appendix

Data Sets

LFW Face Database

Labeled Faces in the Wild

Features -
13,233 Faces
5,749 People
1,680 People with 2+ Images

"..Not enough to provide evidence that a particular software has been thoroughly tested"

Challenges & Limitations

Limitations of the data: no people to test our one-shot identification

Initial classifier is pre-trained on large face dataset

Our Methodology

Further train the model via **"One-Shot Learning"** on a dataset of readily available participants

Integrate our one-shot and pre-training with a train-test split for each

Test on a custom-made dataset that contains a single image of a test participant to be found by the drone

Motivations