# **K-VQG:** Knowledge-based Visual Question Generation

Name: Ammar Hatiya

Supervisor: Dr. H. Davoudi

# Introduction

## What is Knowledge-based Visual Question Generation (K-VQG)?

- Generating questions from images, that can be answered by looking at the image.

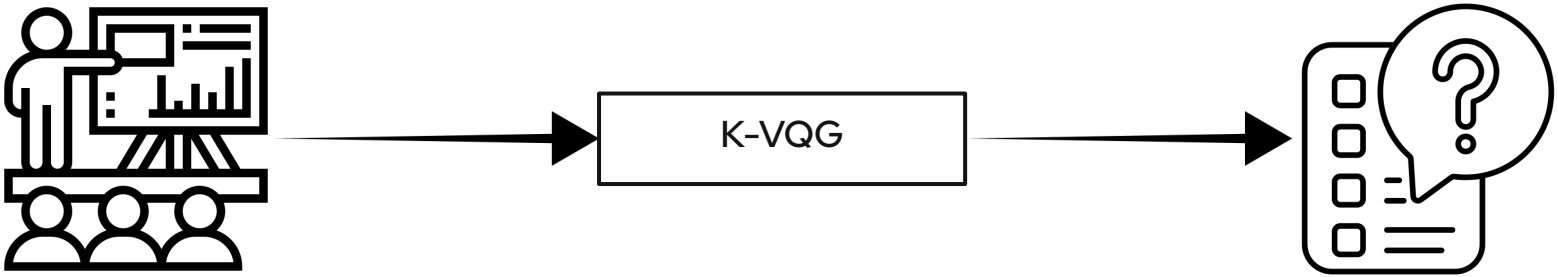
| Question | Knowledge-based? |
|---|---|
| **Q**: What color is the vehicle in front of the building?<br>**A**: Silver | ✔ |
| **Q**: How fast can the vehicle go from 0 to 100 km/h?<br>**A**: 6.4 seconds | ✖ |

# Motivation

## Why is it important?

- K-VQG acts as a **bridge** between visual information and natural language, fostering interdisciplinary research.

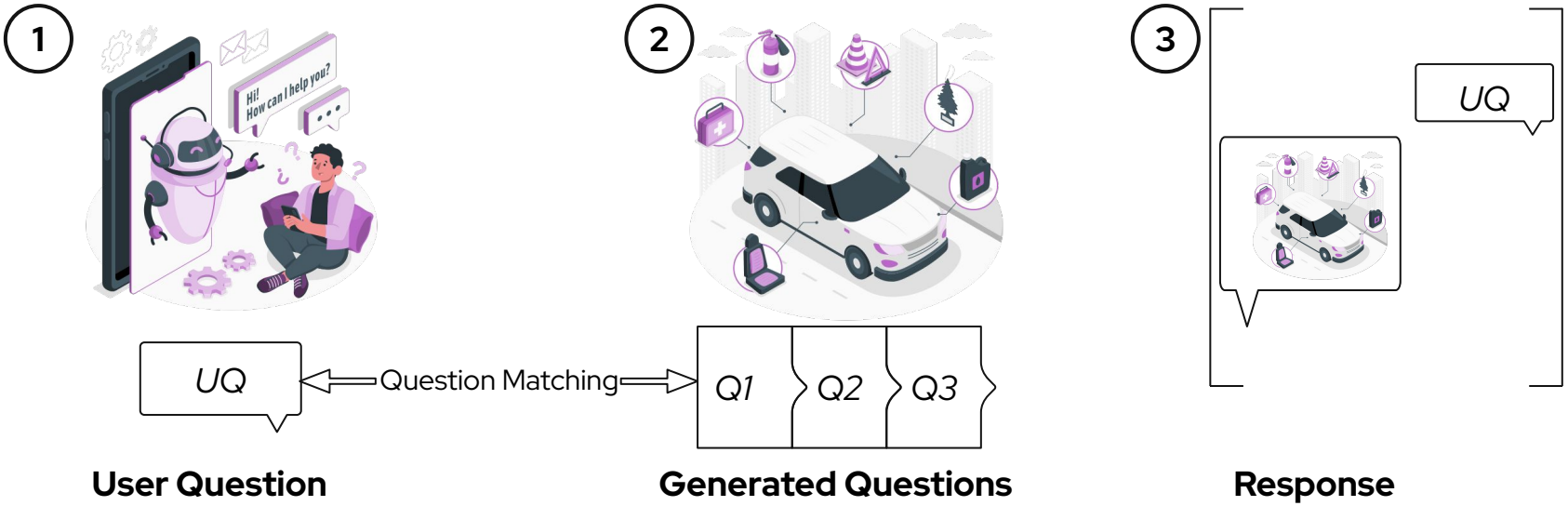
## Potential Application: Educational Content Generation



# Motivation Cont.

## Potential Application: Dialog System for Car Manuals

- Users could ask the system questions related to their car



# Research Objective

## Develop methods for K-VQG based on:

- Large Multimodal Models (GPT-4 Vision)

  Multiple modes of communication or information processing (e.g. Visual, Linguistic)

- Sequence-to-sequence models integrated with Semantic Role Labels (SRLs)

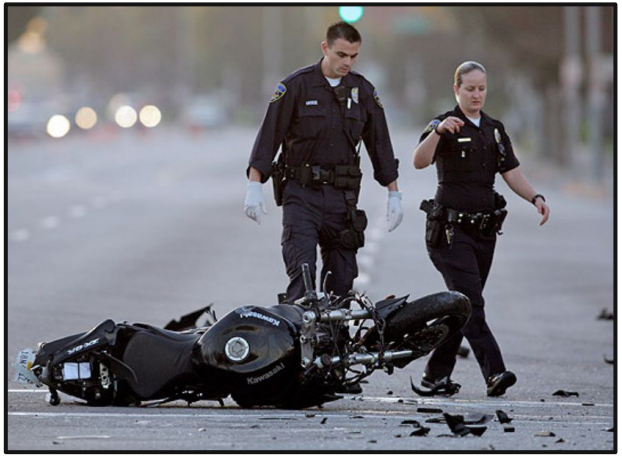
  Linguistic labels that identify the function of words or phrases within a sentence.

## Perform comparative analysis between methods

# Literature Review

## Generating Natural Questions About an Image (2016) [2]

- System should ask ***natural*** and engaging questions about a given image.
- Natural questions: Questions about what is inferred, rather than literal.
- These questions are typically **not** knowledged-based.



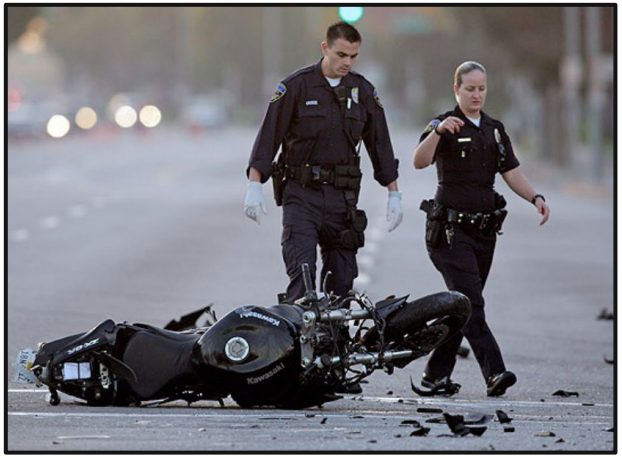
[2]

**Q**: Was anyone injured in the crash?

# Literature Review

## Visual Question Generation for Class Acquisition of Unknown Objects (2018) [3]

- Purpose: Method for generating questions specifically about objects that have not been previously learned
- Unlike knowledge-based question generation, which typically focuses on generating questions that can be answered by observing the content of an image



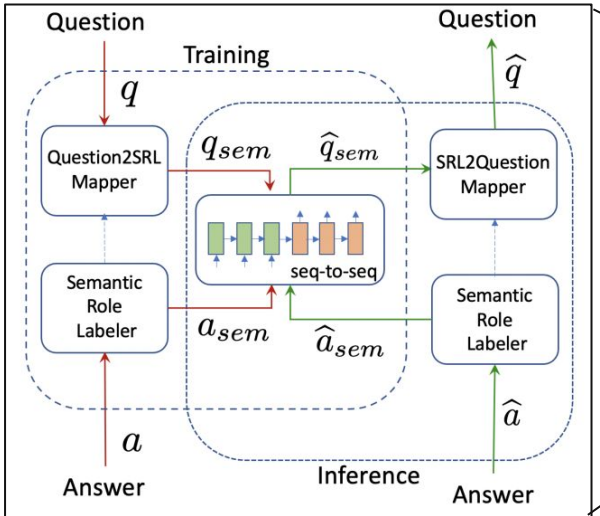
[2]

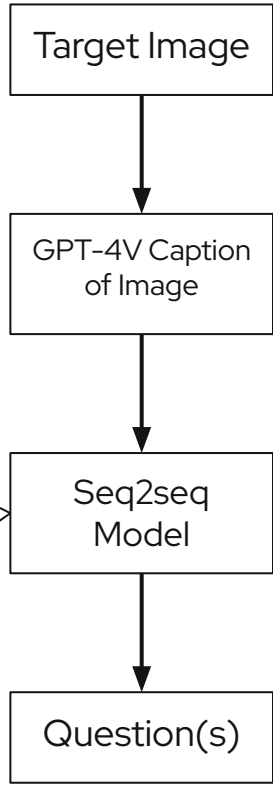
**Q:** What is the object on the ground in front of the police officer?

# Methodology

## Seq2seq (SRL) Based Method[4]

- Used GPT-4V to generate exhaustive description of the images
- Feed this description as input to **Seq2seq** model empowered by Semantic Role Labels, which generates questions from text

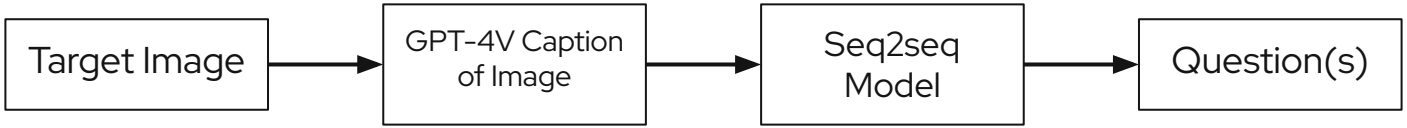


# Methodology

## Seq2seq (SRL) Based Method[4]

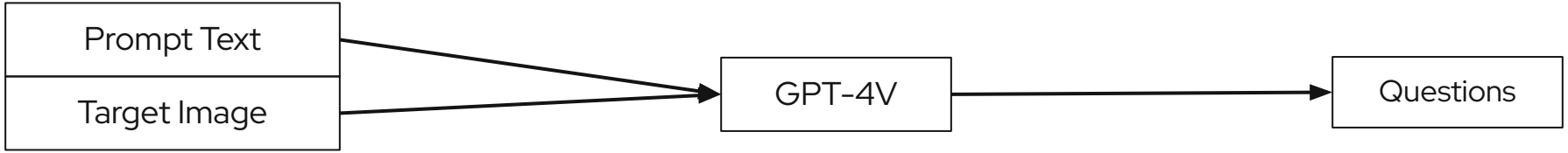
- Used GPT-4V to generate exhaustive description of the images
- Feed this description as input to **Seq2seq** model empowered by Semantic Role Labels, which generates questions from text



# Methodology Cont.

## Vanilla Prompt Engineering Method

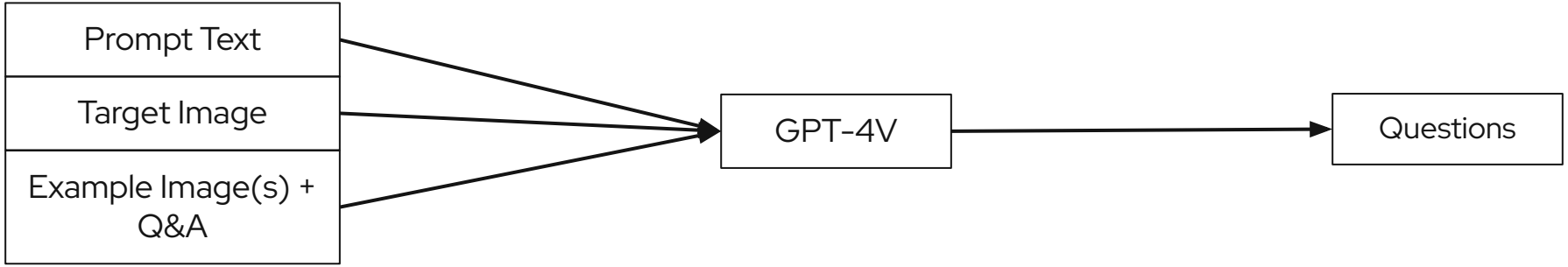
- Given an image, prompts GPT-4V to generate 5 questions and answers from the image, that can be answered by looking at the image
- Serves as **baseline** prompt, to understand the effects of different context additions



# Methodology Cont.

## 1-Shot, 2-shot Prompt Engineering Methods

- When presented with an image, GPT-4V is prompted similarly to the "Vanilla" Prompt method, with the addition of one or two **examples**, respectively.
- Example consists of: Image + 5 knowledge-based questions & answers



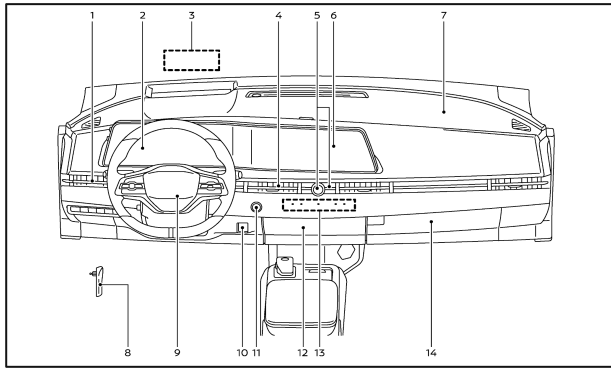
# Methodology Cont.

## 1-Shot + Categorization Prompt Engineering Method

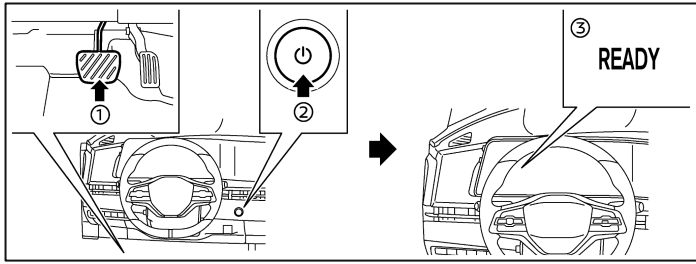
- Categorization of image
  - Can be automated (~70% accuracy)

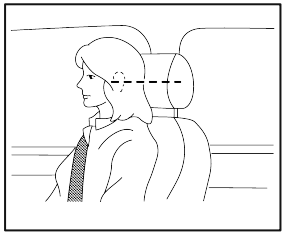
- Label each image as 1 of 3 types:
  - **Item Dictionary**
  - **Procedural**
  - **Other**

(Item Dictionary)

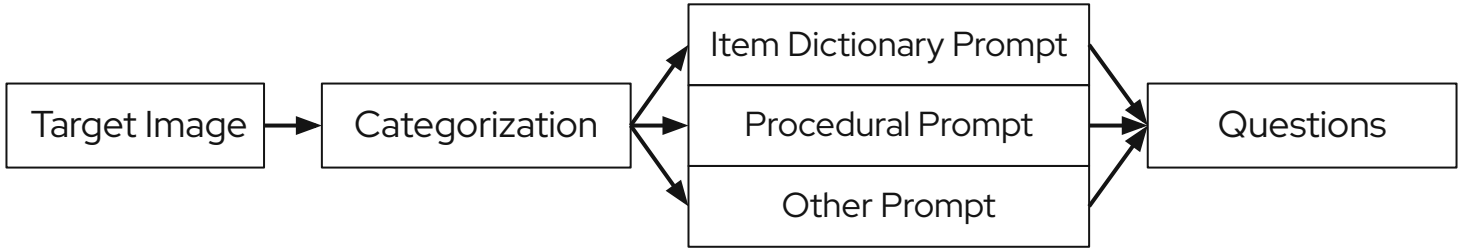



(Procedural)



(Other)



# Setup

## Dataset:

- 2023 Nissan Ariya (EV) Owner's Manual
- Consists of **451** Images

## Creation of Ground Truth:

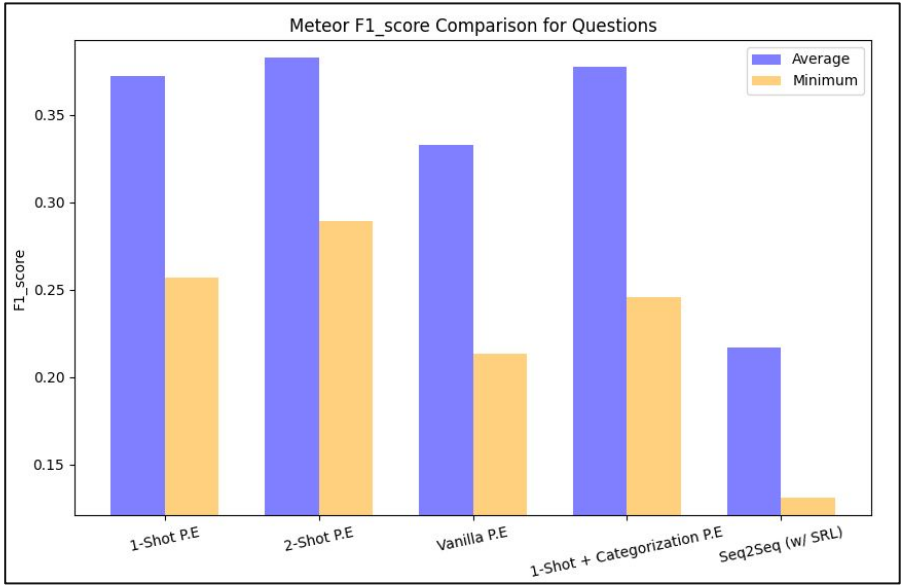
- Curated **132** Knowledge-based questions & answers from **17** images

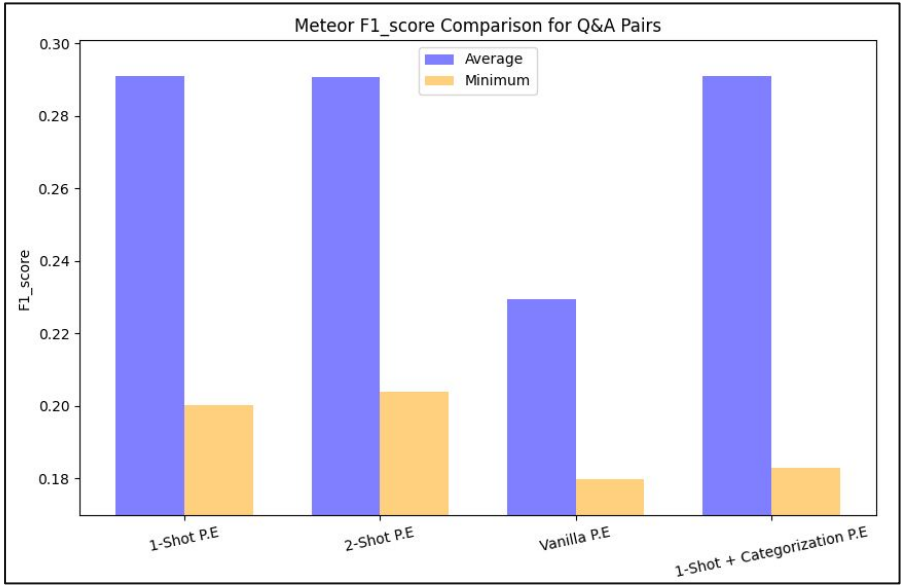

[5]

# Results Cont.

## Metric: METEOR Similarity Score

- Considers both lexical (word) and semantic (meaning) similarities
- Is **sensitive** to the order of words in a question

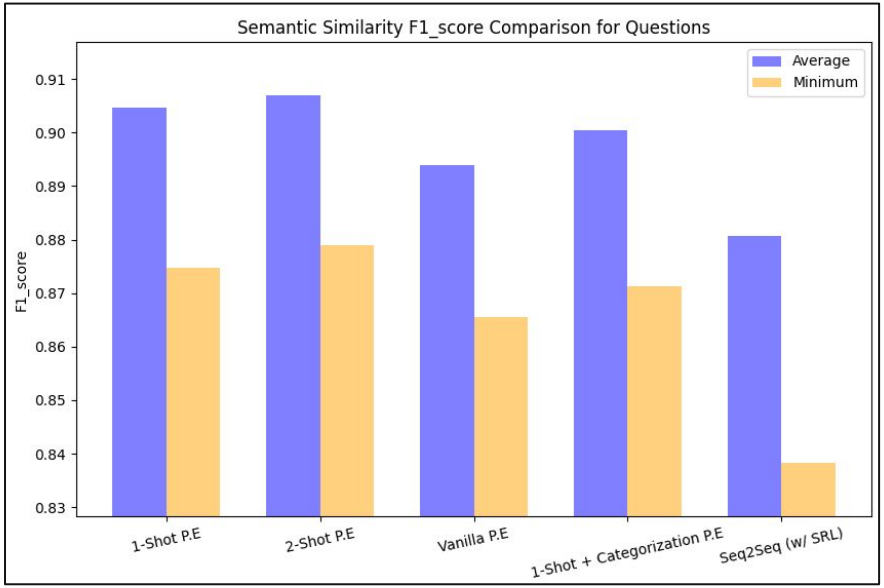


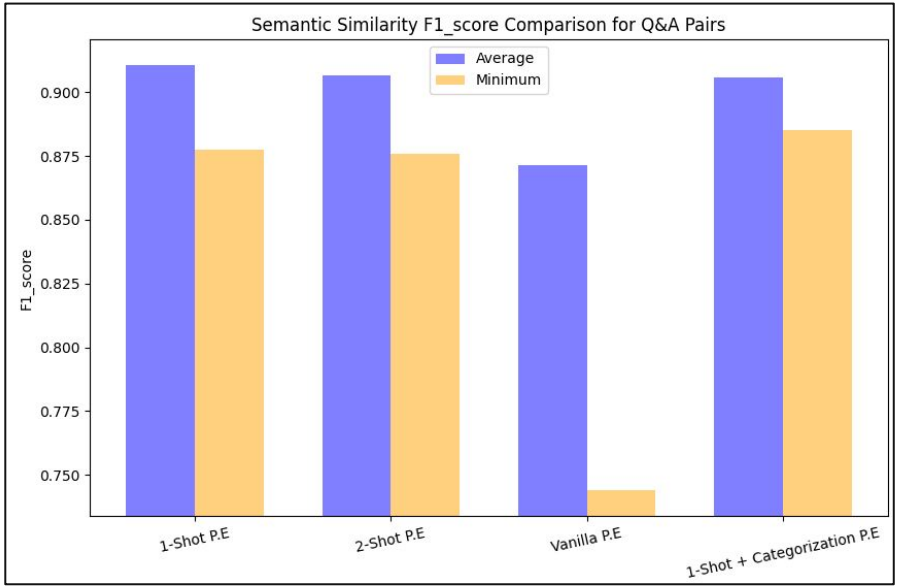



# Results

## Metric: Semantic Similarity Score (spacy)

- Determined by comparing word embeddings, semantic vectors
- **Insensitive** to the order of words in a question



# Conclusion: Findings & Implications

## Method Rankings (for Questions)

1. 2-Shot Prompt Engineering
2. 1-Shot Prompt Engineering
3. 1-Shot + Categorization Prompt Engineering
4. Vanilla Prompt Engineering
5. Seq2seq w/ SRL

GPT-4V demonstrates optimal performance in K-VQG when supplied with **examples** (few-shot learning) and afforded the flexibility to adapt to them.

# Conclusion: Future Work

- Functional Applications:
  - Learning assessments/checks
  - Dialog Systems

- Fine-tuning GPT-4 Vision OR Use different Large Multimodal Models
  - **Unavailable**: Latest model that can be fine-tuned: gpt-4-0613 (*experimental*)
  - Other LMMs: Apple's MM1, Google's Gemeni

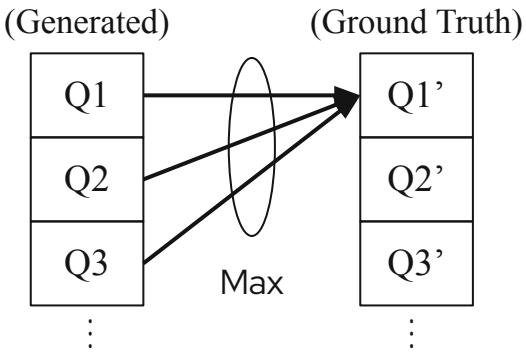
- Data Collection
  - Bigger Dataset (Potential "**Web scraping**")

# Questions?

# Precision, Recall, F1 Score

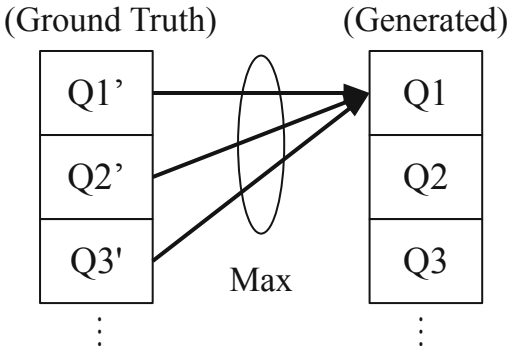
**For an Image X, Precision, Recall, and F1 Score are defined as follows:**

## Precision




Precision = Mean of all Maximums (for each Ground Truth Question)

## Recall



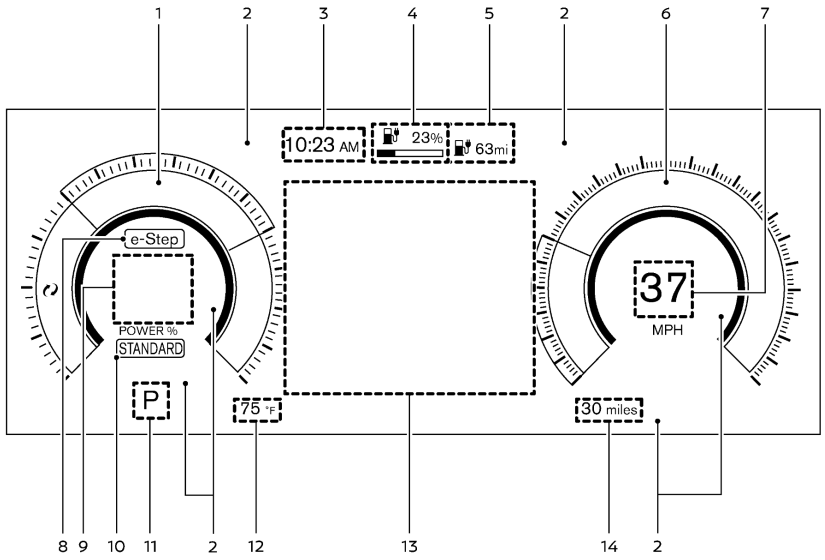

Recall = Mean of all Maximums (for each Generated Question)

## F1 Score

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 Score = Harmonic Mean of Precision and Recall

# Example Results




| Method | Question |
|---|---|
| Seq2seq w/ SRL | What is the second half-circle gauge? |
| Vanilla Prompt Engineering | What time is displayed in the center of the image? |
| 1-Shot Prompt Engineering | What does the 'P' symbol inside a circle on the car dashboard indicate? |
| 2-Shot Prompt Engineering | Is the vehicle in motion according to the image? |
| 1-Shot + Categorization Prompt Engineering | Where is the Fuel Gauge located as per this diagram? |

# Thanks!

If you have any questions:

ammar.hatiya@ontariotechu.net

https://www.linkedin.com/in/ammar-hatiya/

# Sources


Photodoto. (n.d.). Images on the Internet. Retrieved from https://www.photodoto.com/images-on-the-internet/

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). Generating Natural Questions about an Image. Papers with Code. Retrieved from https://paperswithcode.com/paper/generating-natural-questions-about-an-image

Chen, M., Huang, X., Xiao, Z., Wei, J., & Hwang, W. (2020). Visual Question Generation for Class. Papers with Code. Retrieved from https://paperswithcode.com/paper/visual-question-generation-for-class

Naeiji, A., An, A., Davoudi, H., Delpisheh, M., Alzghool, M. (2023). Question Generation Using Sequence-to-Sequence Model with Semantic Role Labels. Paper to be presented at the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), Dubrovnik, Croatia, May 2-6, 2023.

Uehara, K., & Harada, T. (2022). K-VQG: Knowledge-aware Visual Question Generation for Common-sense Acquisition. arXiv preprint arXiv:2203.07890.

Nissan Ariya Concept.jpg [Photograph]. (2024, April 17). Retrieved from https://en.wikipedia.org/wiki/Nissan_Ariya