

# **Multimodal Learning for Visual Question Answering using World Knowledge**

Mohammed Bin Ali Alhaj

MSc in Artificial Intelligence  
The University of Bath  
2024

This dissertation may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

# Multimodal Learning for Visual Question Answering using World Knowledge

Submitted by: Mohammed Bin Ali Alhaj

## Copyright

Attention is drawn to the fact that copyright of this dissertation rests with its author. The Intellectual Property Rights of the products produced as part of the project belong to the author unless otherwise specified below, in accordance with the University of Bath's policy on intellectual property (see [https://www.bath.ac.uk/publications/university-ordinances/attachments/Ordinances\\_1\\_October\\_2020.pdf](https://www.bath.ac.uk/publications/university-ordinances/attachments/Ordinances_1_October_2020.pdf)).

This copy of the dissertation has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the dissertation and no information derived from it may be published without the prior written consent of the author.

## Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Master of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

## Abstract

Navigating the frontier of the Visual Turing Test, this research delves into multimodal learning to bridge the gap between visual perception and linguistic interpretation, a foundational challenge in artificial intelligence. It scrutinizes the integration of visual cognition and external knowledge, emphasizing the pivotal role of the Transformer model in enhancing language processing and supporting complex multimodal tasks.

This research explores the task of Knowledge-Based Visual Question Answering (KB-VQA), it examines the influence of Pre-Trained Large Language Models (PT-LLMs) and Pre-Trained Multimodal Models (PT-LMMs), which have transformed the machine learning landscape by utilizing expansive, pre-trained knowledge repositories to tackle complex tasks, thereby enhancing KB-VQA systems.

An examination of existing Knowledge-Based Visual Question Answering (KB-VQA) methodologies led to a refined approach that converts visual content into the linguistic domain, creating detailed captions and object enumerations. This process leverages the implicit knowledge and inferential capabilities of PT-LLMs. The research refines the fine-tuning of PT-LLMs by integrating specialized tokens, enhancing the models' ability to interpret visual contexts. The research also reviews current image representation techniques and knowledge sources, advocating for the utilization of implicit knowledge in PT-LLMs, especially for tasks that do not require specialized expertise.

Rigorous ablation experiments conducted to assess the impact of various visual context elements on model performance, with a particular focus on the importance of image descriptions generated during the captioning phase. The study includes a comprehensive analysis of major KB-VQA datasets, specifically the OK-VQA corpus, and critically evaluates the metrics used, incorporating semantic evaluation with GPT-4 to align the assessment with practical application needs.

The evaluation results underscore the developed model's competent and competitive performance. It achieves a VQA score of 63.57% under syntactic evaluation and excels with an Exact Match (EM) score of 68.36%. Further, semantic evaluations yield even more impressive outcomes, with VQA and EM scores of 71.09% and 72.55%, respectively. These results demonstrate that the model effectively applies reasoning over the visual context and successfully retrieves the necessary knowledge to answer visual questions.

# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Visual Question Answering (VQA)	2
1.2. Knowledge-Based Visual Question Answering (KB-VQA)	3
<b>2. Literature &amp; Technology Survey</b>	<b>4</b>
2.1. Evolution of Machine Learning	4
2.2. Transformer	5
2.2.1. Attention Mechanism	5
2.2.2. Transformer Architecture	8
2.3. Pre-Trained Large Language Models (PT-LLMs):	10
2.4. Pre-Trained Large Multimodal Models (PT-LMMs):	12
2.5. Knowledge-Based Visual Question Answering (KB-VQA):	14
<b>3. Methodology</b>	<b>18</b>
3.1. Design	18
3.1.1. Image to Language Transformation	19
3.1.2. Pre-Trained Large Language Model (PT-LLM)	21
3.1.3. Prompt Engineering Module	22
3.2. Dataset	23
3.2.1. Overview of Datasets	23
3.2.2. OK-VQA Dataset Analysis	24
3.3. Implementation	26
3.3.1. Component Models	26
3.3.2. Fine-Tuning	26
3.3.3. Hardware & Environment	27
3.3.4. Interactive Model Access on Huggingface Space	28
<b>4. Evaluation and Results</b>	<b>29</b>
4.1. Evaluation Metrics	29
4.2. Evaluation Process	31
4.2.1. Syntactic Evaluation	31
4.2.2. Semantic Evaluation	32
4.3. Main Results	34
4.4. Comparative Results	35
4.5. Qualitative Analysis	36
4.6. Ablation Study	38
<b>5. Ethical Considerations</b>	<b>39</b>
5.1. Limitations & Broader Impact	39
5.2. Reproducibility	39
<b>6. Conclusion &amp; Future Work</b>	<b>41</b>
6.1. Conclusion	41
6.2. Future Work	42
<b>Bibliography</b>	<b>45</b>

- 7. Appendix.....52**
- A. Design ..... 52
  - A.1. YOLOv5 Detectable Object Classes.....52
  - A.2. Token Distribution for Fine-tuning Data .....53
  - A.3. Default LLaMA-2 System Prompt .....53
  - A.4. KB-VQA System Prompt.....54
  - A.5. Comparison of Captioning Models .....54
- B. Implementation ..... 55
  - B.1. Overall Implementation Flow .....55
  - B.2. Component Models Configurations and Hyperparameters .....56
  - B.3. Fine-tuning configurations and Hyperparameters .....56
  - B.4. Fine-tuning Results .....57
- C. Evaluation..... 57
  - C.1. Token Count for Evaluation Data (DETIC vs YOLOv5) .....57
  - C.2. Ablation Study Scores (Graph).....58
  - C.3. Ablation Study Results per Question Category .....59
  - C.4. Additional Evaluation Samples .....61
- D. LLaMA-2 License by Meta..... 62

# List of Figures

Figure 1-1: Examples from VQA dataset v1.-----2

Figure 1-2 Human process to handle KB-VQA.-----3

Figure 1-3 KB-VQA sample.-----3

Figure 2-1 Attention Mechanism.-----6

Figure 2-2 Transformer architecture with single layer of encoder and decoder.-----8

Figure 2-3 Multi-Head Attention.-----9

Figure 3-1 KB-VQA Model Architecture.-----19

Figure 3-2 Object detection and BBOX.-----20

Figure 3-3 Questions distribution over knowledge categories.-----25

Figure 3-4 OK-VQA question keyword distribution.-----25

Figure 3-5 Fine-tuning data structure.-----27

Figure 4-1 Token Counts vs VQA Score.-----37

Figure 6-1 Blueprint design for Language-Vision embeddings alignment for Multimodal learning and Instruction-Following.-----42

Figure 7-1 Token count distribution for the finetuning data before and after removing samples with more than 1024 tokens.-----53

Figure 7-2 Comparison between various captioning models demonstrating InstructBLIP superiority.-----54

Figure 7-3 Overall flow of the project.-----55

Figure 7-4 Fine-tuning learning curves.-----57

Figure 7-5 Token Count for Evaluation Data (DETIc vs YOLOv5)-----57

Figure 7-6 Ablation study results.-----58

Figure 7-7 Additional evaluation samples.-----61

# List of Tables

<i>Table 1-1 Computer vision sub-tasks required to be solved by VQA.</i>	2
<i>Table 2-1 Summary of KB-VQA Methods.</i>	17
<i>Table 3-1 Default prompt template for LLaMA-2 Chat Model.</i>	22
<i>Table 3-2 Customized prompt template for LLaMA-2 Chat Model.</i>	23
<i>Table 3-3 KB-VQA datasets.</i>	24
<i>Table 3-4 OK-VQA dataset characteristics.</i>	25
<i>Table 4-1 GPT-4 API settings for semantic evaluation.</i>	34
<i>Table 4-2 Main results of the KB-VQA system compared to existing methods.</i>	34
<i>Table 4-3 Qualitative visualization of the results.</i>	36
<i>Table 4-4 Ablation experiments for KB-VQA model components.</i>	38
<i>Table 7-1 Component models configurations and hyperparameters.</i>	56
<i>Table 7-2 Fine-tuning configurations and hyperparameters.</i>	56
<i>Table 7-3 Syntactic evaluation results for the ablation study per question category.</i>	59
<i>Table 7-4 Semantic evaluation results for the ablation study per question category.</i>	60

## Acknowledgements

I am profoundly grateful for the support and guidance I have received throughout the course of my dissertation. I would like to extend my deepest appreciation to the following individuals:

To my supervisor, **Dr. Andreas Theophilou**, whose expertise, and insightful guidance have been instrumental in the completion of this research. Your mentorship has not only profoundly shaped my work but also my future endeavours in the field of Artificial Intelligence.

Special mention must be made of my mentors at the University of Bath—**Dr. Ben Ralph, Dr. Hongping Cai, and Dr. Nadejda Roubtsova**. The wealth of knowledge and insights I have gained from you has been indispensable. Your unwavering dedication to academic excellence and steadfast support have been crucial in navigating my academic journey.

My colleagues deserve equal gratitude, for their camaraderie and collaborative spirit have not only made this journey feasible but also deeply enjoyable. The shared experiences and the challenges we have overcome together have been integral to my personal and professional growth.

Lastly, my heartfelt thanks are extended to my family, whose unyielding love and encouragement have been my steadfast anchor. Your belief in my abilities has consistently inspired me and bolstered my strength throughout this process.

This dissertation is not merely a reflection of my individual efforts but stands as a testament to the collective support and wisdom of each individual mentioned above. I am honoured and privileged to be part of such a supportive and enriching academic community.

# Chapter 1

## 1. Introduction

Our perception of the world is inherently multimodal, blending various sensory inputs to create a rich tapestry of experiences. Sight, sound, and touch converge, transforming our environment into a vibrant narrative where our senses actively participate in shaping our understanding. We seamlessly integrate these senses to navigate and interpret our surroundings, relying on vision to perceive objects, auditory cues to detect sounds, and tactile sensations to recognize textures.

The term ‘modality’ broadly refers to the ways in which phenomena manifest and are perceived, closely associated with sensory channels that underpin our primary modes of communication and sensation, notably vision and hearing. In the field of Artificial Intelligence (AI), a problem or dataset that incorporates a variety of such modalities is described as multimodal (Baltrusaitis, Ahuja and Morency, 2018).

The potential of multimodal learning spans a multitude of tasks, each highlighting its versatility and depth. For example: Visual Question Answering (VQA) (Antol et al., 2015) leverages both visual and textual cues to respond to queries about images, while Image Captioning (Vinyals et al., 2015) generates descriptive text for visual content. Text-to-Image Retrieval (Lin et al., 2014) and Text-to-Image Generation (Reed et al., 2016) explore the relationship between linguistic concepts and visual representations, enriching our understanding of both modalities. Additionally, Audio-Visual Speech Recognition (Kim et al., 2013 ; cited in Baltrusaitis, Ahuja and Morency, 2018) and Audio-Visual Emotion Recognition (Yugas et al., 1989 ; cited in Baltrusaitis, Ahuja and Morency, 2018) combine auditory and visual cues to enhance communication and emotional understanding, while Video Analysis (Xu, Zhu and Clifton, 2023) synthesizes temporal and spatial data to interpret complex visual content.

While images and text are both interpreted visually, they represent distinct modalities with unique processing demands. Images are processed holistically, providing spatial information and context instantaneously. Conversely, text is a symbolic, sequential medium that requires abstract cognitive functions for decoding linguistic structures. This distinction underscores the challenges and methodologies necessary for integrating these modalities within AI.

By emulating the human ability to integrate sensory inputs, multimodal learning equips machines with enhanced capabilities to interpret and interact with their environments. For instance, an AI-powered navigation robot utilizes multimodal sensors—integrating visual, auditory, and tactile inputs—to navigate its surroundings effectively. This comprehensive sensory integration enhances the robot's performance and adaptability, showcasing the practical benefits of multimodal learning in sophisticated AI systems.

## 1.1. Visual Question Answering (VQA)

Since its inception by Alan Turing in 1950, the **Turing Test** (Turing, 1950) has been a fundamental benchmark for evaluating machine intelligence against human standards. As technology evolves, so too must the criteria for assessing AI. The **Visual Turing Test** (Geman et al., 2015) represents a modern extension that includes visual cognition within the scope of AI evaluation. At the forefront of this advancement is **Visual Question Answering (VQA)** (Antol et al., 2015), a field that challenges AI systems to perceive, comprehend, and articulate insights about visual inputs in natural language. This progression reflects the complex interplay between perception and cognition that characterizes human intelligence, positioning VQA as a crucial metric for gauging AI’s ability to emulate human-like understanding.

Mature VQA systems hold transformative potential across various domains. In robotics, VQA systems can enhance autonomous decision-making by enabling robots to interpret and respond to visual cues. In medical imaging and diagnosis, VQA systems can assist healthcare professionals by accurately interpreting complex medical images and providing insightful answers to diagnostic questions, thereby enhancing both the speed and accuracy of medical assessments. In manufacturing, VQA systems can optimize quality control processes by enabling automated systems to identify defects and ensure product consistency with minimal human intervention. These advancements underscore the importance of developing robust VQA capabilities, as they push the boundaries of the Visual Turing Test and bring us closer to achieving true human-like AI cognition.

Although manifestation of the VQA task has appeared prior to 2015 as an intersection between Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation (de Faria et al., 2023), the term “Visual Question Answering” (VQA) was coined by Antol et al. (2015), where they proposed the task of free-form and open-ended VQA, defining it as a challenge where a computer is given an image and a natural language question about that image, and it must provide an accurate natural language answer, to support that, they created the first large-scale dataset to evaluate the ability of VQA methods, some examples are shown in [Figure 1-1](#).

CV Task	Representative VQA Question
Object recognition	What is in the image?
Object detection	Are there any elephants in the picture?
Attribute classification	What color is the dog?
Scene classification	Is it raining?
Counting	How many people are there in the
Activity recognition	Is the child crying?
Spatial relationships among objects	What is between the cat and the TV?
common-sense reasoning	Does this person have a perfect vision?

Table 1-1 Computer vision sub-tasks required to be solved by VQA.  
Source: (Manmadhan and Kooor, 2020).



Figure 1-1: Examples from VQA dataset v1.  
Source: (Antol et al., 2015).

Unlike other vision-language tasks, VQA requires many CV sub-tasks to be solved in the process, some of which are summarized by (Manmadhan and Kooor, 2020) in [Table 1-1](#). These VQA tasks often do not require external factual knowledge and only in rare cases require common-sense reasoning (Wu et al., 2022; Reichman et al., 2023). Furthermore, VQA models cannot derive additional knowledge from existing VQA datasets should a question require it (Wu et al., 2022), therefore Knowledge-Based Visual Question Answering has been proposed (Marino et al., 2019).

## 1.2. Knowledge-Based Visual Question Answering (KB-VQA)

In traditional VQA task, the machine is expected to answer a question about an image where the answer requires no more factual information other than what is in the image. For example, when asked ‘How many TVs are there in this room?’ in reference to an image of a living room, the VQA model can provide the correct answer without needing external factual information. However, in reality, humans combine visual observation and logical reasoning with external knowledge when answering questions as illustrated in Figure 1-2.

Consider the image-question pair example in Figure 1-3; the model performing all downstream vision tasks on this image would never be able to know what colour the bus should be without external knowledge linking school buses to the colour yellow.

Knowledge-Based Visual Question Answering (KB-VQA) is a relatively new extension to VQA ( Marino et al., 2019; Reichman et al., 2023) with datasets representing a knowledge-based VQA task where the visual question cannot be answered without external knowledge (Marino et al., 2019), where the essence of this task is centred around knowledge acquisition and integration with the visual contents of the image.

Knowledge sources used for KB-VQA can be arguably categorized into four categories (Lymperaïou and Stamou, 2023) that are not mutually exclusive:

1. **Implicit Knowledge:** Non-symbolic knowledge stored within machine learning models, like the weights of pre-trained neural networks, derived from extensive linguistic and visual data during pre-training. This unstructured knowledge base enhances the model's generic understanding and adaptability to various tasks (Radford et al., 2019; Brown et al., 2020). However, its "black box" nature raises concerns about interpretability and perpetuation of biases or errors from pre-training data, potentially limiting applicability in certain scenarios.
2. **Explicit Knowledge:** Structured information directly accessible and interpretable by a KB-VQA system, often in machine and human-readable formats like ConceptNet (Speer, Chin and Havasi, 2017). It effectively fills gaps in transfer learning addressing unseen concepts and relationships during training, improving understanding of novel concepts and reducing computational demands associated with pre-training (Lymperaïou and Stamou, 2023). However, data collection and curation efforts, especially in specialized domains, are substantial. Aligning and integrating different knowledge graphs can also pose challenges and potentially limit their practical benefits (Ilievski et al., 2021).
3. **Web-Crawled Knowledge:** Gathered from the public internet, structured or unstructured, such as Wikipedia. A significant challenge lies in validating data quality, which can adversely affect the model efficacy (Lymperaïou and Stamou, 2023). Methods for autonomously ensuring data quality are necessary but time-consuming. While offering transparency, explicitness of reasoning in web-sourced knowledge is less pronounced compared to structured knowledge graphs.
4. **Internal Knowledge:** This is the knowledge that the model has learned from the data itself (image-question pair), from visual to the textual embeddings. This type of knowledge is the extent of the vanilla VQA models.

This research investigates the integration of multimodal learning frameworks with advanced machine learning techniques to enhance Knowledge-Based Visual Question Answering systems. By leveraging pre-trained models and incorporating sophisticated inferential mechanisms, the study aims to bridge the gap between visual data interpretation and linguistic analysis, providing a deeper understanding of both the technical challenges and potential solutions within this field.

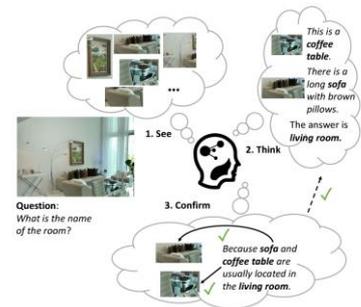


Figure 1-2 Human process to handle KB-VQA. Source: (Chen et al., 2023).



Q: What colour would this be if it were transporting students?

A: Yellow

Figure 1-3 KB-VQA sample. Source: (Marino et al., 2019).

## Chapter 2

# 2. Literature & Technology Survey

### 2.1. Evolution of Machine Learning

Machine Learning (ML) has evolved significantly over the years, driven by innovations and breakthroughs that have shaped its trajectory. The journey from its early beginnings to the emergence of Transformers has been a captivating one, with each phase contributing to the field's progress.

In its nascent stage, ML emerged as a concept in the mid-20<sup>th</sup> century, with Alan Turing's seminal work that laid the foundation for artificial intelligence. Turing's introduction of the Turing Test in 1950 sought to evaluate a machine's ability to exhibit intelligent behaviour, a concept that would become a cornerstone of ML (Turing, 1950).

The 1960s and 1970s saw the rise of symbolic AI and expert systems, which utilized rule-based reasoning. However, these systems faced limitations in scalability and handling uncertainty. The Perceptron (Rosenblatt, 1958), an early neural network model, was introduced but soon encountered the XOR problem, which underscored the limitations of linear models and led to a temporary decline in neural network research (Minsky and Papert, 1969).

The 1980s marked a resurgence in neural network research, spurred by the rediscovery of the backpropagation algorithm. This algorithm allowed for the efficient training of multi-layer neural networks, establishing the connectionist paradigm and emphasizing the distributed and parallel nature of neural computation (Rumelhart, Hinton and Williams, 1986).

The 1990s introduced Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) as powerful tools for binary classification tasks, showcasing exceptional proficiency in areas such as image and speech recognition. Concurrently, this era marked the formalization and widespread adoption of Reinforcement Learning (RL), a paradigm shift eloquently described by Sutton and Barto (1998) which enabled computational models to learn and adapt based on their actions and associated rewards. Furthermore, the application of machine learning to Natural Language Processing (NLP) was significantly advanced, as evidenced in the works of Manning and Schiitze (1999), particularly in tasks like part-of-speech tagging, parsing, and information extraction, thereby enriching the field's capability in understanding and processing human language.

The 2000s represented a seminal era in the progression of ML, ensemble methods such as Random Forests (Breiman, 2001) emerged as a predominant force in ML, exhibiting superior performance across a spectrum of tasks including classification, regression and feature selection. The latter part of the 2000s marked the inception of Big Data era, propelled by an exponential increase in digital data generation.

The 2010s was marked by a significant shift towards deep learning, particularly with the rise of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (LeCun, Bengio and Hinton, 2015). The ImageNet challenge in 2012 became a pivotal moment when a deep CNN, AlexNet (Krizhevsky, Sutskever and Hinton, 2012), drastically outperformed traditional ML. This event catalyzed deep learning research, with CNNs

becoming a staple in computer vision applications. RNN variants like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) gained prominence in NLP, powering chatbots, translation services, and sentiment analysis (Sutskever, Vinyals and Le, 2014). However, the true revolution in NLP arrived with the introduction of the Transformer architecture (Vaswani et al., 2017), originally introduced for machine translation tasks. Transformers marked a departure from recurrent structures, introducing attention mechanisms and self-attention. This breakthrough enabled machines to capture intricate relationships within data more effectively and transformed sequence-to-sequence tasks (Vaswani et al., 2017).

## 2.2. Transformer

The significance of Transformers lies in their ability to capture context and dependencies in data efficiently. The attention mechanisms allow models to weigh the importance of different parts of the input sequence, making them highly effective in tasks involving sequential data, such as language processing and time series analysis.

### 2.2.1. Attention Mechanism

The concept of attention in neural networks, inspired by the cognitive ability to focus selectively on certain aspects of the environment while ignoring others, represents a fundamental shift in the way artificial intelligence systems process information (Itti and Koch, 2001; Mnih, Heess and Graves, 2014). In the human brain, attention allows for the efficient processing of the vast array of sensory inputs by focusing on the most relevant pieces of information. Similarly, in neural networks, attention mechanisms enable the model to dynamically prioritize certain parts of the input data over others, thereby enhancing the efficiency and effectiveness of the processing (Vaswani et al., 2017).

The roots of attention mechanisms can be traced to cognitive science, where attention is recognized as a crucial component of human perception, memory, and decision-making (Posner, 1980). This selective focus capability of the human brain has been a source of inspiration for developing computational models that can mimic this aspect of human cognition.

The early implementation of attention mechanisms in neural networks was proposed by Graves (2013) for tasks such as handwriting recognition and synthesis. This marked the first step in applying the concept of selective focus in computational models.

Another important development in the application of attention mechanisms to neural machine translation was made by Bahdanau, Cho and Bengio (2014) who proposed a novel alignment model that learns to align and translate jointly. Unlike the previous models that encoded the whole input sentence into a single fixed-length vector (Graves, 2013), Bahdanau, Cho and Bengio (2014) introduced the idea of creating context vectors dynamically for each output timestep, effectively addressing the challenge of encoding entire input sequences into fixed-length vectors, a limitation in earlier seq2seq models. This attention mechanism was a breakthrough in handling long-range dependencies in complex sentence structures, significantly improving the capabilities of machine translation models.

The most substantial advancement in attention mechanisms was achieved by Vaswani et al. (2017) with the introduction of the Transformer model, which incorporated a novel structure known as the scaled dot-product attention. This architecture marked a departure from recurrent layers, instead processing the entire input sequence in parallel, substantially improving efficiency and model performance, especially in handling long-range dependencies within the data. The Transformer has since become the foundation for a multitude of state-of-the-art NLP architectures, revolutionizing the field with its innovative approach to sequence modelling.

#### 2.2.1.1. Queries, Keys, and Values ( $q, k, v$ ):

The attention mechanism stands as a significant innovation, facilitating nuanced data processing. Central to this mechanism are three components: queries, keys, and values, each playing a distinct role in how the model processes and interprets sequences.

**Queries ( $q$ ):** In the technical framework of attention mechanisms, queries represent the current element or aspect of data that the model focuses on (Bahdanau, Cho and Bengio, 2014; Vaswani et al., 2017). Analogous to a conscious decision in cognitive processing, a query in this context is akin to a focal point in a sequence, such as a specific word or phrase in a sentence that the model aims to interpret or generate (UI Abideen, 2023).

**Keys ( $k$ ):** Serve as indicators or markers within the data, used to assess the relevance or importance of different parts of the input sequence in relation to the query (Bahdanau, Cho and Bengio, 2014; Vaswani et al., 2017). They can be compared to subconscious cues in human cognition, which, although not always in the forefront of conscious thought, significantly influence attention and interpretation (UI Abideen, 2023). In computational terms, keys are derived from the input data and are used to calculate attention scores, reflecting their relevance to the query (Bahdanau, Cho and Bengio, 2014; Vaswani et al., 2017).

**Values ( $v$ ):** Represent the actual content or information contained in the input data. They are conceptually similar to the information stored within the subconscious mind, which becomes accessible and pertinent when triggered by relevant cues. In the mechanism of attention, values are weighted in accordance with the attention scores derived from the interaction between queries and keys. The aggregation of these weighted values forms the resultant output of the attention step, yielding a representation that is contextually enriched and emphasizes the input aspects deemed most significant (Bahdanau, Cho and Bengio, 2014; Vaswani et al., 2017).

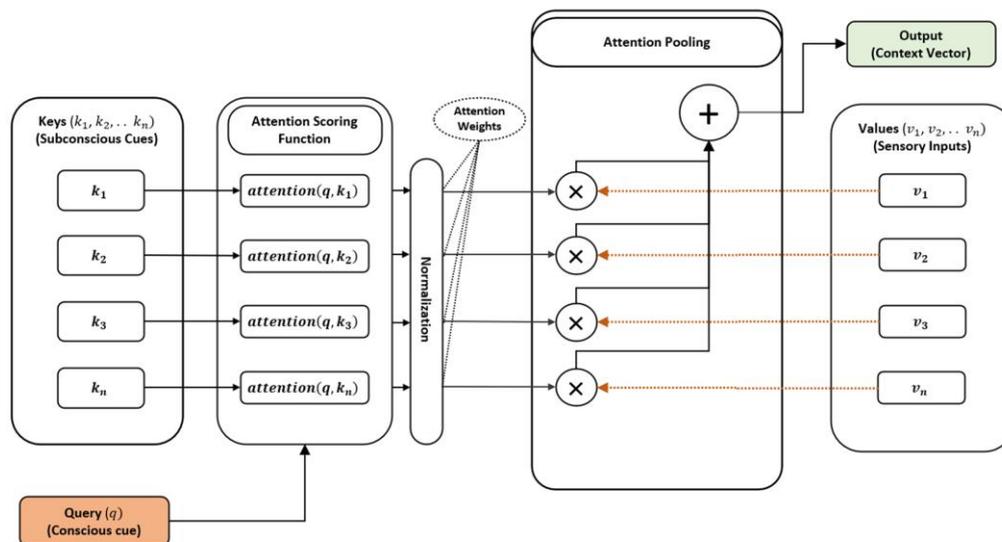


Figure 2-1 Attention Mechanism.

As shown in Figure 2-1, the attention mechanism operates by using queries to interact with keys, determining relevance and thereby assigning weights to the values. The model then aggregates these weighted values to form a contextually enriched output, which incorporates the most relevant information from the input as determined by the queries and keys. This process allows the model to focus on specific parts of the input data, enhancing its ability to process and interpret large and complex data effectively.

### 2.2.1.2. Attention Scoring Functions

A crucial aspect of the attention mechanism is the Attention Scoring Function, which plays a pivotal role in determining how the model assigns relevance to different parts of the input data. This function calculates scores that are used to generate the attention weights, directly influencing the formation of the context vector (output of the attention mechanism). The scoring function evaluates the compatibility or alignment between a query ( $q$ ) and each key ( $k$ ) in the input data. This evaluation results in a set of scores, each representing the relevance of a corresponding value ( $v$ ) in the input sequence to the query ( $q$ ). There are several types of attention scoring functions that have been used since the introduction of attention mechanism in neural network, most common of which include:

- 1. Additive/Concatenative Scoring:** Introduced by Bahdanau, Cho and Bengio (2014), this approach involves concatenating the query and key then feeding them into a feed-forward neural network to produce a score as shown in Equation 2-1. This method allows for the scoring of sequences of different lengths and dimensions.

$$attention(q, k) = w_v^T \tanh(w_q q + w_k k) \in R \quad \text{Equation 2-1}$$

Where  $q$  is the query,  $k$  is the key and  $w_q, w_k$  are learnable parameters.

- 2. Dot-Product Scoring:** Popularized by Luong, Pham and Manning (2015), this method calculates the score as the dot product of the query and key as shown in Equation 2-2. It's a simpler approach but requires the query and key to be of the same dimension.

$$attention(q, k) = q^T k \quad \text{Equation 2-2}$$

Where  $q$  is the query and  $k$  is the key.

- 3. Scaled Dot-Product Scoring:** Central to the Transformer model by (Vaswani et al., 2017), this approach scales the dot product by the inverse square root of the dimension of the key vectors as shown in Equation 2-3. This operation mitigates the potential escalation of SoftMax function inputs to excessively large magnitudes, a scenario that can precipitate gradient instability during the training phase, manifesting as vanishing or exploding gradients. The implementation of this scaling factor is instrumental in maintaining the dot products within a numerically feasible range, thereby ensuring a more robust and efficient training process for the model.

$$attention(q, k) = \frac{qk^T}{\sqrt{d_k}} \quad \text{Equation 2-3}$$

Where  $q$  is the query,  $k$  is the key and  $d_k$  is the key vector dimension.

**Normalization:** Once the scores are calculated, they are normalized, typically using a SoftMax function (Bahdanau, Cho and Bengio, 2014; Vaswani et al., 2017), to form a probability distribution known as attention weights. The SoftMax function guarantees that the total of all attention scores equals one, enabling their interpretation as probabilities or weights (Ul Abideen, 2023). These weights determine how much each part of the input contributes to the context vector (output of the attention mechanism (Bahdanau, Cho and Bengio, 2014; Vaswani et al., 2017; Ul Abideen, 2023)).

### 2.2.1.3. Attention Pooling

Attention pooling plays a pivotal role following the assignment of attention weights. This process involves the multiplication of each value in the input sequence ( $v$ ) by its corresponding attention weight, a step crucial in scaling the values based on their assigned importance (Vaswani et al., 2017). The culmination of attention pooling is the weighted summation of these scaled values (Bahdanau, Cho and Bengio, 2014). This summation effectively aggregates the information from the entire input sequence, with the weights directing focus to the most relevant parts. The outcome of this aggregation is the context vector, a comprehensive representation that encapsulates the most significant information from the input as determined by the model's attention mechanism (Bahdanau, Cho and Bengio, 2014; Vaswani et al., 2017). This context vector serves as a crucial element in subsequent processing steps, embodying the distilled essence of the input data for informed decision-making or output generation (Ul Abideen, 2023).

## 2.2.2. Transformer Architecture

The introduction of the Transformer model by (Vaswani et al., 2017) marked a paradigm shift in the approach to sequence modelling, particularly in the field of natural language processing. Prior to the Transformer, Recurrent Neural Networks (RNNs), often in tandem with attention mechanisms, were the standard for handling sequential data (Bahdanau, Cho and Bengio, 2014; Luong, Pham and Manning, 2015). However, the Transformer architecture innovated by eschewing RNNs in favour of an entirely attention-based approach, thereby addressing some of the inherent limitations of recurrent models.

Most leading neural models designed for sequence transduction adopt an encoder-decoder framework (Bahdanau, Cho and Bengio, 2014). In this structure, the encoder transforms a sequence of symbol representations  $(x_1, x_2, \dots, x_n)$  into a series of continuous representations  $z = (z_1, z_2, \dots, z_n)$ . Utilizing  $z$ , the decoder sequentially produces an output sequence  $(y_1, y_2, \dots, y_m)$  of symbols, generating each element individually. The model operates in an autoregressive manner, incorporating previously generated symbols as additional input to produce subsequent ones (Vaswani et al., 2017).

Employing this foundational structure and as shown in Figure 2-2, the Transformer model utilizes layers of stacked self-attention and pointwise, fully connected components within both its encoder and decoder blocks (Vaswani et al., 2017). The decoder then takes this condensed representation and endeavours to construct the target sentence from it (Bansal, 2023).

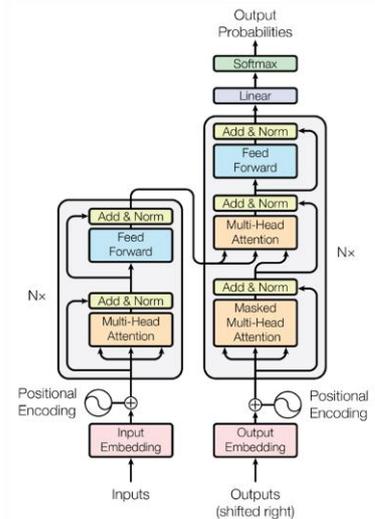


Figure 2-2 Transformer architecture with single layer of encoder and decoder.  
Source: (Vaswani et al., 2017)

### 2.2.2.1. Transformer Encoder

The block on the left side of the Transformer architecture shown in Figure 2-2 is the encoder. It consists of 6 identical layers, each layer has two sub-layers: a multi-head self-attention mechanism, and a position-wise fully connected feed-forward network. It employs a residual connection around each sub-layer followed by layer normalization. All sub-layers and embedding layers produce outputs of dimension ( $d_{model} = 512$ ). The encoder's function is to map the input sequences into continuous representations that hold the learned information including positional encodings (Vaswani et al., 2017).

### 2.2.2.2. Transformer Decoder

The block on the right side of the Transformer architecture shown in Figure 2-2 is the decoder. It consists of 6 identical layers, it has 3 sub-layers: a masked-multi-head self-attention mechanism and a position-wise fully connected feed-forward network, and between these two sub-layers, the encoder has a third sub-layer, which performs multi-head attention over the output of the encoder stack, with residual connections around each sub-layer. The decoder interprets the encoded input and autoregressively generates the output sequence, predicting each token based on the encoder's output and the previously generated tokens in the sequence.

### 2.2.2.3. Multi-Head Self-Attention

Self-Attention represents a significant shift from traditional vanilla attention mechanisms (Vaswani et al., 2017). Unlike conventional methods that map queries to keys from different sequences, as commonly seen in encoder-decoder models (Bahdanau, Cho and Bengio, 2014), self-attention in the Transformer evaluates relationships within the same sequence. This involves transforming each token in the input sequence into a trio of vectors: queries, keys, and values. The self-attention mechanism then calculates attention scores by taking the scaled dot product of the query with all keys, even with itself, which are then passed through a SoftMax function to derive the weights. These attention weights are used to produce a weighted sum of the values, resulting in an

output vector that encapsulates information from the entire sequence, for each element (Vaswani et al., 2017; Voita et al., 2019).

Further enhancing this approach, the Transformer employs multi-head self-attention. This mechanism multiplies the self-attention process across several independent heads ( $h$ ) as shown in Figure 2-3, enabling parallel computation of different representational aspects of the sequence (Vaswani et al., 2017; Voita et al., 2019). The outputs of various attention heads are merged and subsequently transformed via linear projections, culminating in an integrated representation that encapsulates information from a multitude of perspectives (Vaswani et al., 2017).

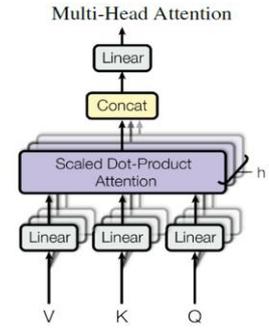


Figure 2-3 Multi-Head Attention.  
Source: (Vaswani et al., 2017).

#### 2.2.2.4. Positional Encoding

Given that self-attention, the core mechanism of the Transformer, is inherently order-invariant and does not differentiate the positions of symbols, positional encoding becomes essential. It explicitly incorporates position information into the input embeddings, addressing this limitation of self-attention (Vaswani et al., 2017).

The implementation of positional encoding in the Transformer utilizes *Sinusoidal* functions with varying frequencies. Each positional encoding vector, which aligns dimensionally with the input embedding vector, is added to the input embedding. This combined vector is then processed by the self-attention layer. The positional encoding vector follows the formulas:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\left(\frac{2i}{d_{model}}\right)}}\right) \quad PE(pos, 2i + 1) = \cos\left(\frac{pos}{10000^{\left(\frac{2i}{d_{model}}\right)}}\right)$$

Equation 2-4

Where ( $pos$ ) is the position, ( $i$ ) is the dimension index, and ( $d_{model}$ ) is the dimension of the input embedding vector (Vaswani et al., 2017).

The benefits of positional encoding for the Transformer model are multifold. Primarily, it enables the model to recognize the order and position of symbols, fostering a more effective learning of data structure and meaning. Additionally, it enhances the model's performance and generalization capabilities by reducing ambiguity in the self-attention mechanism. Importantly, it also circumvents the need for recurrent or convolutional layers to encode positional information, which suffer from vanishing gradients, and are often computationally intensive and challenging to parallelize (Vaswani et al., 2017).

#### 2.2.2.5. Residual Connections, Layer Normalization and Feed Forward Network

Residual connections, also known as skip connections (Orhan and Pitkow, 2018; Kim, 2019), allow the output of one layer to bypass one or more subsequent layers and be added directly to the output of a later layer. This approach facilitates smoother gradient flow across layers during backpropagation (Drozdal et al., 2016). Layer normalization, on the other hand, is applied immediately after residual connections. It normalizes the output across the features for each data sample, stabilizing the learning process by maintaining a consistent scale of activations throughout the network (Xu et al., 2019).

The position-wise feed-forward network within the Transformer architecture constitutes a critical sublayer applied independently and uniformly to each position in both the input and output sequences. Structurally, this network encompasses two linear transformations with a Rectified Linear Unit (ReLU) activation function situated between them. Its primary function is to process each position in the sequence individually, applying the same transformation across all positions irrespective of their respective contexts. This uniform application across positions serves to enhance the local features of the input and output representations, introducing a necessary element of non-linearity into the overall model architecture. Notably, while the position-wise feed-forward network maintains consistent parameters across all positions within a given layer, these parameters differ

between layers in both the encoder and the decoder. This design enables the network to provide distinct transformations at each layer, thereby enriching the model's capacity to capture and integrate complex features and dependencies present in the data (Vaswani et al., 2017).

#### 2.2.2.6. Masked Multi-Head Self-Attention

The masked multi-head self-attention component is a distinctive feature of the decoder. Its primary function is to ensure that the prediction for a specific token is conditioned only on preceding tokens, maintaining the autoregressive nature of the decoder. This is achieved by masking the model's ability to attend to future tokens in the sequence during training. In practical terms, the masking mechanism modifies the attention scores such that each token can only attend to tokens that precede it, effectively preventing information flow from future tokens. This is done by adding  $(-\infty)$  to the attention scores for the future tokens, which makes them zero after applying the SoftMax function.

#### 2.2.2.7. Training and Inference

During training, the Transformer's encoder receives an input sequence, which is first converted into embeddings and subsequently augmented with positional encodings to preserve the sequential information. This composite data passes through multiple encoder layers, each applying self-attention and position-wise feed-forward networks, progressively refining the input's representation. Simultaneously, the decoder is fed with a right-shifted version of the output sequence, again transformed into embeddings, and enriched with positional encodings. The decoder's layers work to predict the next token in the sequence. The masked self-attention ensures that each position in the decoder can only attend to earlier positions in the output sequence, thereby preserving the autoregressive property. The encoder-decoder attention mechanism (middle sub-layer) allows the decoder to focus on relevant parts of the input sequence, utilizing the full context provided by the encoder (Vaswani et al., 2017; Alammari, 2018; Bansal, 2023).

During inference, the process is inherently iterative: starting with an initial token (like a start-of-sequence token), the decoder generates one token at a time, each time reprocessing the sequence of generated tokens and using the encoder's output to predict the next token, until an end-of-sequence token is generated. This sequential generation, informed by the rich context encoded by the encoder and the sequential dependencies learned by the decoder, allows the Transformer to produce coherent and contextually relevant output sequences (Vaswani et al., 2017; Alammari, 2018; Bansal, 2023).

### 2.3. Pre-Trained Large Language Models (PT-LLMs):

Language stands as a cornerstone of human capability, enabling expression and communication. This ability emerges in the early stages of childhood and continually refines throughout one's life. Unlike humans, machines inherently lack the proficiency to understand and communicate using human language. To bridge this gap, they require the integration of advanced artificial intelligence (AI) algorithms. The quest to empower machines with the ability to read, write, and converse in a manner akin to human interaction has been a focal point of research for many years. This pursuit not only challenges our understanding of language itself but also pushes the boundaries of technological innovation in the realm of AI. Central to these efforts is the domain of Language Modelling (LM). Technically, LM provides the foundation to enhance the linguistic intelligence of machines, serving as a critical approach to replicate and potentially surpass human-like language comprehension and generation.

**A Language Model (LM)** is a probabilistic machine learning model used to predict the next word in a sequence. LM aims to model the likelihood-of sequences to predict the probabilities of future (or missing) tokens. It captures the statistical properties of the text in a given language and can be utilized for a range of tasks such as speech recognition, machine translation, and text generation (Zhao et al., 2023). Language models can be either

generative or discriminative, depending on whether they model the joint probability  $P(w_1, w_2 \dots w_n)$  or the conditional probability  $P(w_n | w_1 \dots w_{(n-1)})$  of the tokens (Ng and Jordan, 2001).

**Large Language Models (LLMs)** refer to LMs that have undergone extensive training on wide-ranging textual data, allowing them to internalize the subtleties and probabilistic structure of natural language (Brown et al., 2020). These models are identified by their expansive architectures, typically containing tens or hundreds of billions of parameters<sup>1</sup>. Through training on vast text corpora, LLMs acquire the ability to discern fine-grained nuances and generate text sequences that are not only coherent but also contextually relevant. The predominant architecture employed in these models is the Transformer-based neural network (Vaswani et al., 2017) which contributes to their significant computational demands during both the training phase and subsequent operations. Despite these demands, LLMs demonstrate an impressive capacity for natural language comprehension and the aptitude to tackle and solve complex linguistic tasks (Zhao et al., 2023).

**Pre-Trained Large Language Model (PT-LLM)** is a Large Language Model that has been previously trained on a large-scale corpora, such as Wikipedia, Common Crawl, or Books Corpus, to learn general linguistic patterns and representations (Mikolov et al., 2013; Pennington, Socher and Manning, 2014; Devlin et al., 2018). These models can then be fine-tuned or adapted for specific downstream tasks, such as text summarization, question answering, or sentiment analysis, by using a smaller amount of task-specific data (Howard and Ruder, 2018; Radford et al., 2019; Liu et al., 2019). The main motivation for using PT-LLMs is to leverage the rich knowledge and semantic information encoded in the pre-trained parameters, which can significantly reduce the data and computational requirements for achieving high performance on various natural language processing (NLP) tasks (Radford et al., 2019; Brown et al., 2020).

Qiu et al. (2020) outline several advantages of utilizing PT-LLM for downstream tasks, including better representations learned from the huge text corpus, better model initialization and therefore better generalization, and leveraging the PT-LLM as a regularization to avoid overfitting on the specific task small data.

**Word2Vec** (Mikolov et al., 2013) and **GloVe** (Global Vectors for Word Representation) (Pennington, Socher and Manning, 2014) are examples of the early attempts to pre-train word embedding models on large text corpora which can be used as inputs for LLMs, focusing on word-level representations that captured syntactic and semantic word relationships. These techniques were foundational and played an important role in various NLP tasks, but they did not account for the context of word usage, which limited their effectiveness for complex language understanding tasks (Xu Han et al., 2021). **ELMo** (Embeddings from Language Models) (Peters et al., 2018) was proposed to capture context-aware word representations by pre-training a bidirectional Long Short-Term Memory (biLSTM) network to generate dynamic word embeddings that reflect the surrounding words, rather than relying on static representations. These context-sensitive embeddings were then tailored to enhance performance on a variety of specialized tasks through further task-specific training.

The seminal development of the Transformer architecture (Vaswani et al., 2017) catalyzed significant advancements in the field of Natural Language Processing (NLP), facilitating the emergence of sophisticated models. **BERT** (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) stands as a pivotal achievement in this domain, employing a bidirectional Transformer network and a novel masked language modelling objective to set new standards across various NLP tasks. The impact of BERT precipitated a wave of innovations, with models such as **RoBERTa** (A Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019), **ELECTRA** (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) (Clark et al., 2020), **ERNIE** (Enhanced Language Representation with Informative Entities) (Zhang et al., 2019), and **DeBERTa** (Decoding-Enhanced BERT with Disentangled Attention) (He et al., 2021) enhancing the foundational design through expanded datasets, optimized training processes, and inventive pre-training objectives, which collectively achieved enhanced performance metrics.

In parallel, the Generative Pre-Trained Transformer, known as **GPT**, introduced by Radford et al. (2018), marked a departure from BERT's bidirectionality, adopting an autoregressive, unidirectional mechanism specifically geared towards text generation. The subsequent iterations, **GPT-2** (Radford et al., 2019) and **GPT-3** (Brown et

<sup>1</sup> While the literature does not uniformly agree on a definitive scale limit for LLMs, the predominant discussions often center around models featuring billions of parameters (Zhao et al., 2023).

al., 2020), chronicled a significant progression, showcasing that increased model dimensions and training data could significantly advance language task capabilities (Brown et al., 2020; Xu Han et al., 2021; Zhao et al., 2023). Notably, GPT-3, with its vast 175 billion parameters, underscored the vast capabilities of PT-LLMs, demonstrating aptitudes across a spectrum of complex tasks, mirroring human-like proficiency in translation and creative writing (Brown et al., 2020; Zhao et al., 2023).

Another generation of PT-LLMs was **T5** (Text-To-Text Transfer Transformer) (Raffel et al., 2019) and **BART** (Lewis et al., 2019) which introduced variations of the Transformer that combined both autoencoding and autoregressive techniques, showing that different pre-training objectives could lead to improvements in certain types of language understanding and generation tasks. These models were pre-trained on a denoising objective, where the model learns to reconstruct the original text from a corrupted version, which helps the model learn from diverse and noisy data sources.

More recently, several new PT-LLMs have emerged, such as **PaLM** (Pathways Language Model ) (Chowdhery et al., 2022) that uses pathways to scale language modelling with the Transformer architecture. Pathways are a new ML system that enables highly efficient training across multiple TPU Pods, reducing the communication overhead and improving the throughput, and it gets better with its smaller multilingual variant **PaLM-2** (Anil et al., 2023) that demonstrated better performance than PaLM with reduced training and inference cost (Naveed et al., 2023). **LLaMA** (Large Language Model Meta AI) (Touvron et al., 2023a) is trained on a large and diverse dataset of 101 languages, and it incorporates human feedback into its learning process, and **LLaMA-2** (Touvron et al., 2023b) is a fine-tuned version of LLaMA, optimized for dialogue use cases with a larger context length and grouped-query attention. LLaMA-2 outperforms open-source chat models on most benchmarks and is designed to be helpful and safe and considered popular in the research community for parameter-efficient and instruction tuning (Naveed et al., 2023). **FALCON** (Fused Attention for Language understanding and CONversation) (Penedo et al., 2023) is an autoencoding model that uses a novel objective called Fused Attention and combines masked language modelling, next sentence prediction, and response prediction to learn from both monologue and dialogue data, it achieves state-of-the-art results on several dialogue tasks. **BLOOM** (Big Language Open-Source Model) (Scao et al., 2023) is an autoregressive model that uses a sparse attention mechanism to handle long-range dependencies and large inputs, it is trained on a massive dataset of 1.6 trillion tokens, including web texts, books, and images.

The evolution of PT-LLMs continues to be a vibrant area of research, with models becoming ever more sophisticated, and their applications increasingly widespread. Each new model builds upon the insights and lessons learned from its predecessors, driving forward the capabilities of machine understanding and generation of human language.

## 2.4. Pre-Trained Large Multimodal Models (PT-LMMs):

While LLMs have exhibited remarkable capabilities in zero or few-shot reasoning across a wide range of NLP tasks, they are limited in their inability to process visual information, as they are designed to interpret only textual data. Simultaneously, significant advancements are being made in large vision foundation models in terms of perception capabilities. In response to this gap, unimodal LLMs and vision models are converging, leading to the emergence of the field of Large Multimodal Models (LMMs). From the perspective of developing Artificial General Intelligence (AGI), LMMs represent a significant advancement over LLMs for several reasons (Yin et al., 2023): LMMs align more closely with human perception, which naturally involves receiving and integrating multisensory inputs that are often complementary and synergistic, potentially making LMMs more intelligent; LMMs offer a more user-friendly interface, as their support for multimodal inputs allows users to interact and communicate with intelligent assistants in a more flexible manner; and LMMs are more versatile in task-solving, being able to support a broader range of tasks beyond the typical NLP tasks manageable by LLMs.

Inspired by the advancement of Transformer-based models in NLP, researcher started to see the potential of large-scale Transformer-based vision models. **Vision Transformers (ViTs)** (Dosovitskiy et al., 2021) are a testament to this potential. Bypassing convolutional layers, they interpret images as non-overlapping patch

sequences that are embedded and processed by Transformer layers achieving superior performance compared to CNNs, especially when trained on large data. **Swin-Transformer (Shifted Window Transformer)** (Liu et al., 2021) is another milestone in computer vision, it uses shifted windows to partition the image into overlapping patches of different scales for representation learning.

Building upon the specialized capabilities of NLP and vision pre-trained models, researchers have started to explore multi-modal pre-trained models designed to understand and generate content across multiple data types (Zhou et al., 2020; Han et al., 2023). These models aim to unify the strengths of NLP and computer vision techniques, delivering more comprehensive and nuanced performance on complex tasks that involve both text and images such as VQA and KB-VQA. One of the early attempts in this direction can be attributed to **VLP (Vision-Language Pre-training)** (Zhou et al., 2020). This model integrates visual and textual modalities using a Transformer-based architecture and has been effective in various cross-modal tasks such as image captioning and visual question answering. Following VLP, **CLIP (Contrastive Language-Image Pretraining)** (Radford et al., 2021) made its mark in 2021. CLIP integrates visual and textual information within a unified Transformer framework. It employs a contrastive loss function that encourages the model to produce similar embeddings for semantically correlated text and image data, thereby achieving remarkable results on zero-shot learning tasks (Zero-Shot learning allows a model to transfer knowledge gained during training to new, unseen scenarios). Also in 2021, **DALL-E** (Ramesh et al., 2021) emerged as a variant of the **GPT-3** model trained to generate high-quality images from textual descriptions. While primarily an image generation model, its training methodology exemplifies how multi-modal learning can facilitate the synthesis of content across different data types. The most recent addition to the roster is **GPT-4** (OpenAI, 2023), which was trained using an extraordinary scale of compute and data. GPT-4 is a large multimodal model that can take image and text inputs and produce text outputs. In addition to its linguistic proficiency, GPT-4 is capable of tackling a wide range of complex and new tasks, encompassing fields like mathematics, computer programming, visual perception, healthcare, legal studies, and psychology, all without requiring specialized prompting.

Recent work like **BLIP (Bootstrapping Language-Image Pre-training)** (J. Li et al., 2022) has demonstrated the ability to perform various multi-modal tasks, including Image-Text retrieval and Image Captioning. BLIP effectively utilizes noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones, achieving promising results on a wide range of vision-language tasks. **BLIP-2** (Li et al., 2023) builds upon this by introducing a new visual-language pre-training paradigm that leverages any combination of pre-trained vision encoder and LLM without having to pre-train the whole architecture end-to-end. This model bridges the modality gap with a lightweight Querying Transformer (Q-Former), enabling state-of-the-art results on multiple visual-language tasks while reducing pre-training costs. **InstructBLIP** (Dai et al., 2023) takes the concept of instruction tuning from language models and applies it to vision-language models. It introduces an instruction-aware Query Transformer, which extracts informative features tailored to the given instruction, leading to state-of-the-art zero-shot performance across various tasks. **LLaVA** (Liu et al., 2023) represents a novel end-to-end trained large multimodal model that combines a vision encoder with an LLM by linearly projecting the visual features into the LLM embedding space, achieving impressive chat capabilities and setting a new state-of-the-art accuracy on Science QA dataset (Lu et al., 2022). It is a cost-efficient approach to building general-purpose multimodal assistants.

Adding to these, **Gemini** (Google, 2023) emerges as a family of highly capable multimodal models that exhibit remarkable capabilities across image, audio, video, and text understanding. The Gemini family consists of Ultra, Pro, and Nano sizes, suitable for applications ranging from complex reasoning tasks to on-device memory-constrained use-cases.

Almost all vision-language pre-trained models utilize a Transformer-based pre-trained Model as a text encoder, but how to learn visual representations based on visual contents is still an open problem (Du et al., 2022).

## 2.5. Knowledge-Based Visual Question Answering (KB-VQA):

(Reichman et al., 2023) has categorized the current approaches to solving KB-VQA task based on the knowledge sources they use:

1. **Knowledge Graphs (Explicit Knowledge):** These are structured databases that store information about the world. They are a popular choice for knowledge sourcing because they are easy to access and interpret by ML models (Gui et al., 2022).
2. **Passage Retrieval (Web-Crawled):** This involves using a search engine to retrieve relevant passages from text documents. The passages are then used to answer the questions (Gao et al., 2022a).
3. **Fine-tuning Vision-Language Models (Implicit Knowledge):** These are pre-trained models that have been trained on a large dataset of images and text. They can be fine-tuned to answer questions about images by using a knowledge graph or other knowledge sources (Ding et al., 2022).
4. **Prompt-Based Large Language Models (e.g., GPT-3) to access the knowledge from the models (Implicit Knowledge):** These models are trained to generate text in response to prompts. The prompts can be designed to guide the models to answer questions about images (Shao et al., 2023).

KB-VQA constitutes a challenging task for AI models, it necessitates the capability to understand and analyze visual and textual data while also utilizing external knowledge to derive precise answers (Marino et al., 2019).

Various research efforts have explored different approaches to tackle the task of KB-VQA, **AHAB**<sup>2</sup> (Wang et al., 2017) was one of the early efforts, it processes the inputs into structured queries and retrieves supporting knowledge from fixed knowledge bases to obtain the answers. Further efforts focused on acquiring the knowledge from open-domain sources such as Wikipedia and ConceptNet (Speer, Chin and Havasi, 2017), then perform joint reasoning over the retrieved knowledge, image, and question to predict the answer. **ConceptBERT** (Gardères et al., 2020) represents an initial foray into an end-to-end transformer-based approach that is fully reliant on open-domain explicit knowledge, where it simultaneously leverages multiple modalities for learning. The process begins with generating representations for each modality separately. Image is represented as a list of objects with their bounding-boxes using Faster R-CNN (Ren et al., 2016), while BERT (Devlin et al., 2018) embeddings are employed for question representation and ConceptNet (Speer, Chin and Havasi, 2017) serves as the source of common-sense knowledge. The model comprises two primary modules: a vision-language module, and a concept-language module based on a bidirectional transformer architecture. These modules interact to form a unified concept-vision-language representation, which is then processed through a classifier to produce the final answer. However, the model struggles with questions that require complex reasoning.

The performance of these methods is often limited, either by the non-existence of required knowledge for that specific question or the irrelevant noisy knowledge that is inevitably retrieved (Shao et al., 2023). In an attempt to better utilize the noisy retrieved knowledge, **MAVEx (Multi-modal Answer Validation using External knowledge)** (Wu et al., 2021) added an answer validation mechanism, but that’s only useful if the knowledge exists!

Inspired by the success of Transformer-based pre-trained models, multiple efforts have focused on using implicit knowledge. **KRISP (Knowledge Reasoning with Implicit and Symbolic rePresentations)** (Marino et al., 2021) uses a transformer-based model and utilizes object regions to learn implicit knowledge stored in BERT (Devlin et al., 2018) as a supplementary knowledge resource to the structured knowledge base. However, the complexity of combining diverse knowledge sources may lead to challenges in maintaining the explicit semantics of symbolic knowledge.

**PICa (Prompting GPT-3 via the use of Image Captioning)** (Yang et al., 2022) combined acquiring the implicit knowledge in KRISP (Marino et al., 2021) and the validation stage in MAVEx (Wu et al., 2021) by unlocking the first use of GPT-3 (Brown et al., 2020) for multimodal tasks using frozen GPT-3 model as an implicit and unstructured knowledge base, PICa first converts the image into a list of tags using Microsoft Azure Tagging API<sup>3</sup> and uses VinVL (Zhang et al., 2021) to generate image captions that GPT-3 can understand, then utilizes

<sup>2</sup> AHAB: the captain in the novel “Moby Dick” (Wang et al., 2017).

<sup>3</sup> Public Azure Tagging & Captioning API: <https://westus.dev.cognitive.microsoft.com/docs/services/computer-vision-v3-2>

GPT-3 to solve the VQA task in a few-shot manner by just providing a few in-context VQA examples. **KAT** (**K**nowledge **A**ugmented **T**ransformer) (Gui et al., 2022) integrates implicit and explicit knowledge in an encoder-decoder architecture, while still jointly reasoning over both knowledge sources during answer generation. KAT generates image tags that are used for explicit knowledge retrieval leveraging contrastive learning between image [CLS]<sup>4</sup> token produced by CLIP (Radford et al., 2021) and knowledge entries representations in the explicit knowledge bases, it also leverages GPT-3 as an implicit knowledge source and treats VQA as an open-ended text generation task. **REVIVE** (**RE**gional **VI**sual Representation for knowledge-based **VI**sual qu**ES**tion answering) (Lin et al., 2022) uses an object detector to locate the objects in the image using GLIP (Grounded Language-Image Pre-training) (L.H. Li et al., 2022), then crops the objects bounding-boxes and uses the cropped region proposals to retrieve different types of external knowledge, then GPT-3 (Brown et al., 2020) is prompted with regional tags, question and a caption generated by VinVL (Zhang et al., 2021) to retrieve the implicit knowledge, the model then integrates both knowledge inputs with the regional visual features into a unified transformer based answering model for final answer generation. **MuKEA** (**MU**ltimodal **K**nowledge **E**xtraction and **A**ccumulation framework) (Ding et al., 2022) accumulates multimodal knowledge from VQA samples, without relying on existing Knowledge-Bases, it represents multimodal knowledge by triplets that correlate visual objects and fact answers with implicit relations. It uses a pre-training and fine-tuning strategy to learn the triplet representations from different views, such as embedding structure, topological relation, and semantic space. **TRiG** (**T**ransform-**R**etrieve-**G**enerate) (Gao et al., 2022b) employs a methodology that converts all visual context into textual representations utilizing a combination of a captioner (Li et al., 2020), an object detector (Han et al., 2021), and Optical Character Recognition (OCR)<sup>5</sup> modules. Subsequently, it conducts downstream tasks exclusively within the linguistic domain. Specifically, it utilizes the textual representation of the visual context to retrieve the top-k knowledge passages from Wikipedia. This retrieval is facilitated by calculating the dot product between the CLS token of the encoded visual context, which is processed by BERT (Devlin et al., 2018), and the embeddings of various knowledge passages. Furthermore, TRiG encodes both the question and the visual context alongside each passage among the top-k selected knowledge passages using the T5 (Raffel et al., 2019) encoder. It then concatenates the embeddings of the tuples (question, visual context, knowledge) and inputs them into the T5 decoder. The decoder, trained autoregressively, is tasked with generating the answer.

**IPVR** (Interactive Prompting Visual Reasoner) (Z. Chen et al., 2023) tries to mimic human process to solving KB-VQA task by introducing three modules: *See – Think – Confirm*, the *See* module detects the objects in the image and translate the image into a global description using Faster R-CNN variant (Xiaotian Han et al., 2021), *Think* module adopts OPT (Zhang et al., 2022) to select semantically relevant visual concepts extracted by the *See* module corresponding to the given task, then transforms them into textual descriptions using BLIP (J. Li et al., 2022), lastly, OPT predicts the answer based on the attended visual context. *Confirm* module is used for rationale, it requires the LLM to continue generating answer’s supporting rationale which is verified using CLIP (Radford et al., 2021). The *Think – Confirm* process iteratively continues until the answer predictions in two consecutive iterations are consistent.

(Shao et al., 2023) argues that all previous approaches that utilized GPT-3 have not fully activated the capacity of the giant model as the provided input information to the model is insufficient and proposes **PROPHET** (**PRO**mPt GPT-3 with answer **HE**uristics for knowledge-based VQA), initially, they modified MCAN-Large (Yu et al., 2019) with a grid-based features extracted from CLIP (Radford et al., 2021) visual encoder instead of the original bottom-up-attention region-based features and a BERT-Large (Devlin et al., 2018) instead of the original LSTM network and used it as a vanilla VQA model that is then trained on a KB-VQA dataset without the aid of external information, following this, two mutually complementary answer heuristics—answer candidates and answer-aware examples—are extracted from the trained model. Lastly, these heuristics are incorporated into the prompt structure to augment GPT-3’s understanding of the task, thereby improving its performance capabilities.

<sup>4</sup>Classification token used in transformer-based models and added to the beginning of the sentence tokens in text and the beginning of the sequence of image patches (tokens) if its image to aggregate the global information from the entire sentence or image (Devlin et al., 2018; Dosovitskiy et al., 2021).

<sup>5</sup> They used off the shelf EasyOCR (<https://github.com/JaidedAI/EasyOCR>).

**LAMOC (Language Model Guided Captioning)** (Du et al., 2023) leverages the strengths of both captioning and language models. It uses Reinforcement Learning to leverage the guidance and feedback of the PT-LLM to tune the captioning model using a probabilistic reward function and a relevance score as the reward signal. In this way, the caption provides context for the PT-LLM, while the PT-LLM's feedback refines the caption to be more informative for answering the question. **Two (Thinking while Observing)** (Si et al., 2023) employs a dual-encoder framework, multimodal encoder that functions as an 'Observer' to encode visual features, and a textual encoder, acting as a 'Thinker,' to encode a diverse array of knowledge resources. This framework also includes an answer decoder, which decodes the latent embeddings from both encoders to generate the final response. The system integrates implicit knowledge sourced from GPT-3 (Brown et al., 2020) and enhances this with implicit multimodal knowledge from OFA (Wang et al., 2022), which has been fine-tuned on the VQAv2 dataset (Goyal et al., 2017). Additionally, it incorporates explicit knowledge extracted from Wikipedia. The methodology involves utilizing the image caption, object list, and the posed question to prompt GPT-3 for knowledge acquisition, while simultaneously prompting OFA with the image features and the question to access implicit multimodal knowledge. All retrieved knowledge is then input into the textual encoder, and concurrently, the image and question are processed by the multimodal encoder. The outputs of these encoders are projected into a unified embedding space, and the combined output is subsequently fed into the answer decoder to produce the final answer.

Recently, **Q&APrompts** (Wang and Ge, 2024) introduced a novel approach by incorporating a pre-trained large multimodal model (PT-LMM) to the system for solving KB-VQA task. It focuses on discovering rich visual clues by mining question-answer pairs in images and using them as prompts for PT-LMM. The process involves three key stages: Visual Question Generation (VQG) Model Training: In this stage, image-answer pairs, and corresponding questions from a VQA training set are used to train a VQG model. This model learns to map an answer with an image to generate relevant questions. Question-Answer Prompts Generation: Using an image tagging model (Y. Zhang et al., 2023), various instances within an image are identified. These tagged image pairs are then fed into the VQG model to generate pertinent questions with the extracted image tags serving as answers. Visual-Language Reasoning: The generated question-answer pairs are encoded as prompts with a visual-aware prompting module. These prompts are then sent to InstructBLIP (Dai et al., 2023) that acts as the PT-LMM to deduce the final answers. Q&APrompts has shown substantial improvements in performance when tested on challenging datasets like OK-VQA (Marino et al., 2019). Similarly, **GeReA (Generate-ReASON)** framework (Ma et al., 2024) extends the application of PT-LMMs by utilizing InstructBLIP (Dai et al., 2023) and LLaVA-1.5 (Liu et al., 2023) demonstrating that these models provide enhanced visual understanding capabilities over PT-LLMs. GeReA operates through two principal stages: initially, the Question-Aware Prompt Caption Generation, where InstructBLIP and LLaVA-1.5 are prompted with key image regions identified using a cross-attention matrix that measures the similarity between image patch features and related question features, supplemented with customized, question-specific manual prompts to ensure the generated captions closely align with the query specifics. Subsequently, in the Question-Aware Prompt Caption Reasoning stage, these captions are integrated into a multimodal reasoning model that combines the caption embeddings with the image-question pair and similar examples, processed through a T5 (Raffel et al., 2019) decoder to construct a robust joint representation for accurate answer derivation. Although GeReA has shown significant improvements on benchmark datasets like OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022), outperforming prior state-of-the-art methods, it's important to note that the framework's reliance on multiple large models necessitates substantial computational resources, which could affect the speed during inference, highlighting a potential limitation in terms of scalability and efficiency in practical applications.

[Table 2-1](#) summarizes the methods that attempted to solve the KB-VQA, along with their main components used for visual representation, knowledge sources, use of PT-LMs within the model, as well as the model's

performance on OK-VQA dataset (Marino et al., 2019). Models that have not been tested on OK-VQA dataset such as AHAB (Wang et al., 2017) are not included.

Model	Image Representation	Captioner	Detector	PTLM <sup>6</sup>	Knowledge Type	Accuracy (%)
<b>ConceptBert</b> (Gardères et al., 2020)	Tags	-	Faster R-CNN (Ren et al., 2016)	BERT (Devlin et al., 2018)	Explicit: ConceptNet	33.66
<b>MAVEx</b> (Wu et al., 2021)	Feature Embeds.	-	-	ViLBERT (Lu et al., 2019)	Explicit: Wikipedia + ConceptNet + Google Images	39.4
<b>KRISP</b> (Marino et al., 2021)	Feature Embeds.	-	-	MMBERT (Khare et al., 2021)	Explicit: DBpedia + ConceptNet + VisualGenome + haspartKB Implicit: BERT	38.9
<b>PICa</b> (Yang et al., 2022)	Caption + Tags	VinVL (Zhang et al., 2021)	Microsoft Azure Tagging API	GPT-3 (175B) (Brown et al., 2020)	Implicit: GPT-3	48
<b>KAT</b> (Gui et al., 2022)	Feature Embeds + Caption	Oscar (Li et al., 2020)	Faster R-CNN (Ren et al., 2016)	GPT-3 (175B) (Brown et al., 2020) T5-Large (Raffel et al., 2019)	Explicit: Wikidata Implicit: GPT-3	54.41
<b>REVIVE</b> (Lin et al., 2022)	Feature Embeds + Caption + Tags	VinVL (Zhang et al., 2021)	GLIP (L.H. Li et al., 2022)	GPT-3 (175B) (Brown et al., 2020)	Implicit: GPT-3 Explicit: Wikidata	58
<b>MuKEA</b> (Ding et al., 2022)	Feature Embeds + Tags	-	Faster R-CNN (Ren et al., 2016)	LXMERT (Tan and Bansal, 2019)	Explicit: Multimodal knowledge from VQA 2.0 and OK-VQA	42.59
<b>TRIG</b> (Gao et al., 2022b)	Caption + Tags + OCR <sup>7</sup>	OSCAR (Li et al., 2020)	SGG (Han et al., 2021)	T5-Large (Raffel et al., 2019)	Explicit: Wikidata	50.5
<b>IPVR</b> (Z. Chen et al., 2023)	Caption + Tags	BLIP (446M) (J. Li et al., 2022)	SGG (Han et al., 2021)	OPT-66B (Zhang et al., 2022)	Implicit: OPT	44.62
<b>PROPHET</b> (Shao et al., 2023)	Feature Embeds + Caption + Tags	VinVL (Zhang et al., 2021)	Faster R-CNN (Ren et al., 2016)	GPT-3 (175B) (Brown et al., 2020)	Implicit: GPT-3	61.1
<b>LAMOC</b> (Du et al., 2023)	RL Guided Captions	BLIP (446M) (J. Li et al., 2022)	-	FLAN-T5-XXL (11B) (Chung et al., 2022)	Implicit: FLAN-T5-XXL	40.31
<b>Two</b> (Si et al., 2023)	Feature Embeds + Caption + Tags + OCR <sup>1</sup>	OFA (Wang et al., 2022)	VinVL (Zhang et al., 2021)	GPT-3 (175B) (Brown et al., 2020) T5-Large (Raffel et al., 2019) OFA (0.93B) (Wang et al., 2022)	Explicit: Wikidata + Multimodal knowledge from VQA 2.0 Implicit: GPT-3 + OFA + T5	58.72
<b>Q&amp;APrompts</b> (Wang and Ge, 2024)	Feature Embeds	-	RAM (Y. Zhang et al., 2023)	InstructBLIP (7B) (Dai et al., 2023)	Implicit: InstructBLIP	64.3
<b>GeReA</b> (Ma et al., 2024)	Feature Embeds + Caption	InstructBLIP (7B) (Dai et al., 2023) LLaVA-1.5 (7B) (Liu et al., 2023)	-	T5-Large (Raffel et al., 2019) InstructBLIP (7B) (Dai et al., 2023) LLaVA-1.5 (7B) (Liu et al., 2023)	Implicit: InstructBLIP LLaVA-1.5	66.5

Table 2-1 Summary of KB-VQA Methods.

Accuracy results are based on VQA-Score<sup>8</sup> metric evaluated on OK-VQA dataset.

<sup>6</sup> PTLM refers to Pre-Trained Large Models including Pre-Trained Large Language or Multimodal Models.

<sup>7</sup> They used off the shelf EasyOCR (<https://github.com/JaidedAI/EasyOCR>).

<sup>8</sup> VQA Score discussed in Section 4.1.

## Chapter 3

# 3. Methodology

Recent advancements in Pre-Trained Large Language Models (PT-LLMs) have significantly enhanced the capabilities of Knowledge-Based Visual Question Answering (KB-VQA) systems. These models facilitate sophisticated reasoning and enable the extraction of implicit knowledge, which is crucial for interpreting complex visual data. However, a fundamental challenge remains: PT-LLMs do not directly process images as they operate within the textual domain. To address this, current methodologies predominantly focus on two approaches. The first involves aligning visual and textual embeddings into a unified embedding space, using methods such as linear projectors (Liu et al., 2023) or attention-based alignment modules (Dai et al., 2023) which enables the resultant model to perform multiple downstream vision tasks, not limited to KB-VQA alone. Further discussion of this approach can be found in [Section 6.2](#). The second approach, language mediation, which is the focus of this research, converts visual contexts into textual representations. Models like PICA (Yang et al., 2022) and KAT (Gui et al., 2022) exemplify this approach by translating visual data into text, allowing the PT-LLM to leverage its extensive pre-trained knowledge to 'understand' the images textually. This method is resource-efficient and directly leverages the PT-LLM's existing capabilities in natural language processing, making it an ideal strategy for the current scope of this research.

### 3.1. Design

The proposed Knowledge-Based Visual Question Answering (KB-VQA) model integrates insights from various prior works (Marino et al., 2021; Gui et al., 2022; Yang et al., 2022; Z. Chen et al., 2023; Du et al., 2023; Si et al., 2023; Tan and Shen, 2023; Wang and Ge, 2024). As illustrated in [Figure 3-1](#), the model operates through a sequential pipeline, beginning with the Image to Language Transformation Module, in this module, the image undergoes simultaneous processing via image captioning and object detection frozen models, aiming to comprehensively capture the visual context and cues. These models, selected for their initial effectiveness, are designed to be pluggable, allowing for easy replacement with more advanced models as new technologies develop, thus ensuring the module remains at the forefront of technological advancement. Following this, the Prompt Engineering Module processes the generated captions and the list of detected objects, along with their bounding boxes and confidence levels, merging these elements with the question at hand utilizing a meticulously crafted prompting template. The pipeline ends with a Fine-tuned Pre-Trained Large Language Model (PT-LLMs), which is responsible for performing reasoning and deriving the required knowledge to formulate an informed response to the question.

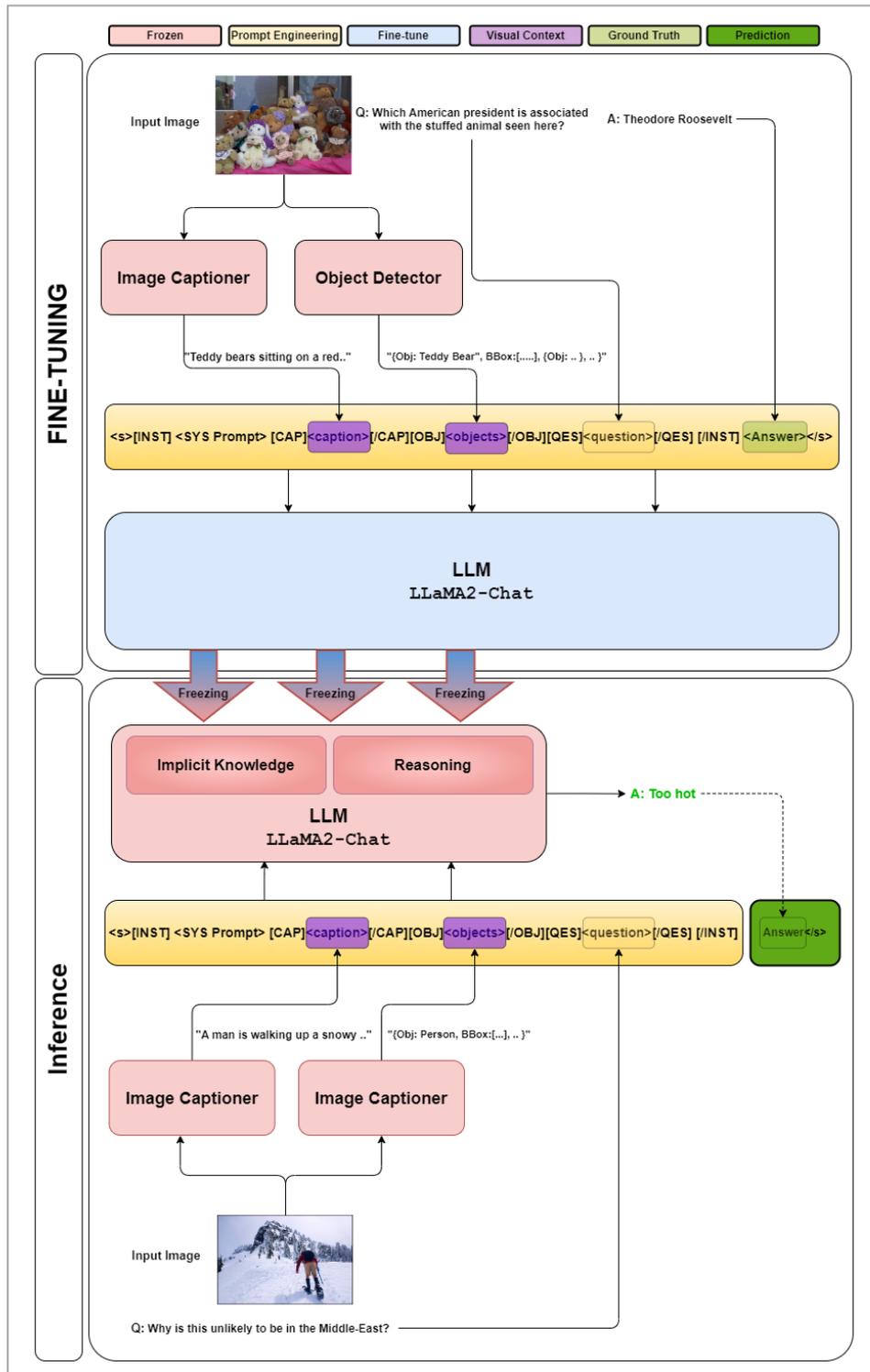


Figure 3-1 KB-VQA Model Architecture.

### 3.1.1. Image to Language Transformation

This stage is pivotal in the process of KB-VQA, during this phase, the wealth of visual information within an image, including the identification of objects, their spatial interrelations, and contextual indicators, is transformed into a coherent linguistic structure. This crucial transformation bridges the divide between unprocessed visual inputs and the following stages of knowledge extraction and logical reasoning. By converting visual elements into a linguistic framework, this phase sets the foundation for the PT-LLM’s ability to not only

access and interpret the required knowledge but also to apply reasoning effectively in responding to intricate visual queries.

### 3.1.1.1. Captioning Module

Most previous works attempting to solve the task of KB-VQA like PICA (Yang et al., 2022) and KAT (Gui et al., 2022) employed OSCAR+ for image captioning and others such as IPVR (Z. Chen et al., 2023) and LAMOC (Du et al., 2023) advanced further by adopting the first version of BLIP family (J. Li et al., 2022). However, these approaches often encountered limitations in providing a comprehensive visual context necessary for the PT-LLM to retrieve the required knowledge and perform reasoning, leading some to implement a closed-loop refinement of captions using Reinforcement Learning (Du et al., 2023). This project leverages the recently released powerful and open-source transformer-based visual instruction tuning model InstructBLIP (Dai et al., 2023) that was trained on 26 datasets and has shown an impressive zero-shot performance in image captioning and other downstream vision tasks (Dai et al., 2023; Ma et al., 2024). InstructBLIP stands out for its instruction-aware Query Transformer, which allows it to generate captions that are contextually relevant to the instructions provided. Its ability to perform exceptionally well in downstream tasks, including complex visual scene understanding and image description, substantiates its suitability as a top-tier image captioner in diverse applications. Additionally, a Huggingface space specifically designed for comparing some state-of-the-art captioning models (Rogge, 2023) was utilized for a quick comparison, where InstructBLIP exhibited superior performance over other models, see A.5 for more details.

### 3.1.1.2. Object Detection Module

To achieve a comprehensive understanding of images, precise detection and localization of objects within these images is essential, a process known as object detection (Felzenszwalb et al., 2008). A crucial aspect of this process is the bounding box (BBOX), which is defined as a rectangular region that encapsulates an object or a designated area of interest within an image. As illustrated in Figure 3-2, this bounding box is demarcated by a set of four coordinates:  $(x_{min}, y_{min})$  represent the top-left corner, and  $(x_{max}, y_{max})$  denote the bottom-right corner of the rectangle (Felzenszwalb et al., 2008). The employment of bounding boxes is a critical methodology for determining the precise location and extent of objects in an image, thereby facilitating spatial awareness of the image contents. By defining the spatial boundaries of each object, bounding boxes enable the model to comprehend the positional relationships between multiple objects in an image, thereby enhancing the model's ability to analyze and interpret complex visual scenes.

The object detection component of the KB-VQA pipelines uses two pretrained detection models:

**YOLOv5 (You Only Look Once)** (Jocher, 2020): This model is pretrained on the COCO dataset (Lin et al., 2014) and is capable of detecting 80 different object classes

**DETIc (DETECTOR with Image Classes)** (Zhou et al., 2022): an encoder-decoder transformer-based object detection and segmentation model pretrained on ImageNet-21K (Deng et al., 2009) and Conceptual Captions (Sharma et al., 2018) datasets, and capable of detecting around 21,000 object classes. This model was primarily utilized to overcome the limitations in the number of detectable objects in YOLOv5.



Figure 3-2 Object detection and BBOX.  
BBOX for object: "person" plotted for demonstration.

The KB-VQA pipeline processes the list of detected objects along with their BBOXes and confidence levels generated by either model in a structured format as shown in below example:

```
{object: tennis racket, bounding box: [61.29, 429.76, 438.9, 633.53], certainty: 86.61%}
{object: person, bounding box: [421.58, 81.35, 1007.64, 940.0], certainty: 86.11%}
{object: sports ball, bounding box: [143.34, 502.11, 195.11, 553.06], certainty: 85.23%}
...
```

### 3.1.2. Pre-Trained Large Language Model (PT-LLM)

Incorporating a Pre-Trained Large Language Model (PT-LLM) within the KB-VQA pipeline significantly enhances its ability to interpret and answer complex visual questions. The profound capabilities of PT-LLMs in natural language understanding and deductive reasoning are crucial for extracting and leveraging implicit knowledge from textually mediated visual data (Zhu et al., 2023).

It can be argued that leveraging the implicit world knowledge stored in PT-LLMs is largely sufficient to handle KB-VQA tasks without the need for explicit knowledge sources or reasoning modules. This assertion is supported by several key points:

1. **Deep Semantic Understanding:** PT-LLMs can interpret detailed captions and recognize relationships among objects in an image, enabling them to answer questions that require more than just visual recognition or syntactic matching (Yang et al., 2022). For example, a PT-LLM might accurately interpret a scene in a photograph where a family is having a picnic under cherry blossoms, recognizing not just the individuals and objects but also the cultural significance and seasonal context, thereby answering questions about the event’s setting or mood.
2. **Leveraging Implicit Knowledge:** Studies have demonstrated that PT-LLMs trained on a large-scale corpus implicitly incorporate substantial knowledge within their parameters. Such Models can then be queried for various types of knowledge, effectively functioning as a knowledge base (KB) (Alkhamissi et al., 2022; Cao et al., 2023). By utilizing structured prompts that encapsulate detailed captions and objects, PT-LLMs leverage this rich, pre-trained knowledge to derive insights that transcend the immediately visible elements of a query. This capability enables them to perform deep, contextual inferences, thereby enhancing their utility in complex interpretative (Gui et al., 2022; Yang et al., 2022; Shao et al., 2023; Wang and Ge, 2024)
3. **Enhanced Deductive Reasoning Capabilities:** The structured information enables PT-LLMs to contextualize the visual content, thereby enhancing their deductive reasoning capabilities. This allows them to derive new facts based on known facts about the scene (Zhu et al., 2023) and make logical connections, which are essential for effectively answering complex questions (Xenos et al., 2023).
4. **Comprehensive Training on Explicit Knowledge Sources:** PT-LLMs undergo extensive training on diverse datasets that inherently include significant sources of explicit knowledge. These sources range from web-crawled data and Wikipedia to public domain books (Touvron et al., 2023b), embedding a wealth of factual and encyclopedic information directly into the model’s parameters. As a result, these models are not only equipped with a broad spectrum of general and applicable knowledge but are also inherently prepared to handle explicit information retrieval and application across a variety of tasks (Brown et al., 2020; Touvron et al., 2023b; Lymperaïou and Stamou, 2023; Zhao et al., 2023; Naveed et al., 2023).
5. **Generalization:** Due to their extensive training, PT-LLMs can generalize well from textual descriptions to unseen contents (Devlin et al., 2018; Brown et al., 2020; Lymperaïou and Stamou, 2023), making them effective at interpreting new images and corresponding questions in KB-VQA tasks.
6. **Adaptability:** With relatively minimal additional training PT-LLMs can be adapted to specific tasks such as KB-VQA through techniques like few-shot learning (Yang et al., 2022) or fine-tuning with task-specific datasets (Touvron et al., 2023b).

The PT-LLM used in this KB-VQA system is the open-source LLaMA-2 (Large Language Model Meta AI-2) (Touvron et al., 2023b), serving as the cognitive center of the KB-VQA pipeline. The LLaMA-2 model family, an advancement over the LLaMA-1 series, is a set of autoregressive decoder-only PT-LLMs and has been pretrained on 2 trillion tokens of data, a 40% increase in training corpus size compared to its predecessor, and features a

context window of 4,096 tokens, up from 2,048 tokens in LLaMA-1. Available in several sizes: 7B, 13B, and 70B parameters. Since its introduction, both LLaMA-1 and LLaMA-2 have garnered significant interest from both academic and industrial sectors (Ruiz, 2023; Zhao et al., 2023; Naveed et al., 2023). These models have demonstrated outstanding performance across various benchmarks, quickly becoming among the most powerful, versatile and favoured open-source language models available (Ruiz, 2023).

As shown in [Figure 3-1](#), this project specifically utilizes the LLaMA-2 Chat, a version that has been fine-tuned from LLaMA-2 using Supervised Fine-Tuning (SFT) and further refined iteratively using Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017), making it optimized for dialogue use cases (Touvron et al., 2023b), and therefore an ideal choice for the KB-VQA system, which involves interactive conversation between the user and the model about an image.

### 3.1.3. Prompt Engineering Module

Prompt Engineering for PT-LLMs is defined by Reynolds and McDonell (2021) as “Programming but in natural language!”, it is the process of designing and optimizing prompts to effectively communicate tasks to PT-LLMs. This discipline involves the strategic formulation of input data (prompts) to elicit specific, desired responses from models that have been trained on vast datasets.

[Table 3-1](#) shows the default prompt template for a single turn conversation and the existing special tokens that LLaMA-2 chat model was trained on.

<b>Default Prompt Template</b>	<pre>&lt;s&gt;[INST] &lt;&lt;SYS&gt;&gt; {system_prompt} &lt;&lt;/SYS&gt;&gt;  {user_message} [/INST] {model_answer}&lt;/s&gt;</pre>	
<b>Description</b>	<code>&lt;s&gt;</code>	Indicates the start of the entire sequence.
	<code>[INST]</code>	Indicates the start of instructions.
	<code>&lt;&lt;SYS&gt;&gt;</code>	Optional: Indicates the start of system prompt.
	<code>{system_prompt}</code>	Optional: A context to guide the model response. (If not provided, the default system prompt is used. (See <a href="#">A.3</a> ))
	<code>&lt;&lt;/SYS&gt;&gt;</code>	Optional: Indicates the end of system prompt.
	<code>{user_message}</code>	The actual user prompt or query.
	<code>[/INST]</code>	Indicates the end of instructions.
	<code>&lt;/s&gt;</code>	Indicates the end of the entire sequence.

*Table 3-1 Default prompt template for LLaMA-2 Chat Model.*

As shown in [Figure 3-1](#), The KB-VQA system is designed to process captions, objects, their bounding boxes, and confidence levels by feeding them to the PT-LLM in a structured linguistic format. To enable the model to understand the given language-mediated visual context and correctly apply reasoning to derive answers, a fully customized prompt template was developed. This includes a tailored system prompt and additional special tokens added to the model’s vocabulary. These tokens indicate the beginnings and ends of captions, details of detected objects, and the questions being asked. The customized system prompt is specifically designed to guide the model on how to effectively utilize these added special tokens. [Table 3-2](#) summarizes the key characteristics of the designed prompt template used in the KB-VQA model. Further details, including the designed system can be found in [Appendix A.4](#).

<b>Customized Prompt Template</b>	<pre>&lt;s&gt;[INST] &lt;&lt;SYS&gt;&gt; {Customized_system_prompt} &lt;&lt;/SYS&gt;&gt;  [CAP] {caption} [/CAP] [OBJ] {objects} [/OBJ] [QES] {question} [/QES] [/INST] {model_answer}&lt;/s&gt;</pre>	
<b>Description</b>	{Customized_system_prompt}	Please see full prompt in <a href="#">Appendix A.4</a>
	[CAP]	Indicates the start of the caption.
	{caption}	The image caption text.
	[/CAP]	Indicates the end of the caption.
	[OBJ]	Indicates the start of the objects list.
	{objects}	The objects along with their BBOXes and confidence levels.
	[/OBJ]	Indicates the end of the objects list.
	[QES]	Indicates the start of the question being asked.
	{question}	The question.
	[/QES]	Indicates the end of the question being asked.

Table 3-2 Customized prompt template for LLaMA-2 Chat Model.

During inference, especially when using DETIC as the object detector with a low confidence threshold, there is a significant chance that the number of detected objects will be high, potentially leading to a prompt token count that surpasses the model's maximum context window of 4,096 tokens. To address this, the model employs a systematic approach: it begins by removing an object from the bottom of the list and then checks if the token count falls within the context window limit. This process of iterative trimming is repeated until the total prompt token count is reduced to below the threshold of 4,000 tokens, maintaining a safe margin of 96 tokens. This ensures that the model operates effectively within its token limit, optimizing performance while adhering to technical constraints.

## 3.2. Dataset

In the field of Knowledge-Based Visual Question Answering (KB-VQA), the creation and refinement of specialized datasets are pivotal. These datasets not only facilitate the training and evaluation of artificial intelligence models but also reflect the diverse and growing complexities that this domain contends with.

### 3.2.1. Overview of Datasets

Multiple datasets have been developed to serve and address the KB-VQA task. Prominent among them are:

- 1. Knowledge-Based VQA (KB-VQA)** (Wang et al., 2015): One of the earliest datasets in this domain, KB-VQA comprises 700 images and 2,402 questions, with each question associated with both an image and a knowledge base (KB). The KB encapsulates facts about the world, including object names, properties, and relationships, aiming to foster models capable of answering questions through reasoning over both the image and the KB.
- 2. Factual VQA (FVQA)** (Wang et al., 2017): This dataset includes 2,190 images and 5,826 questions, accompanied by a knowledge base containing 193,449 facts. The FVQA's questions are predominantly factual and less open-ended compared to those in KB-VQA, offering a different challenge in knowledge-based reasoning.
- 3. Outside-Knowledge VQA (OK-VQA)** (Marino et al., 2019): OK-VQA poses a more demanding challenge than KB-VQA, featuring an open-ended knowledge base that can be updated during model training. This dataset contains 14,055 questions and 14,031 images. Questions are carefully curated to ensure they require reasoning beyond the image content alone.

**4. Augmented OK-VQA (A-OKVQA)** (Schwenk et al., 2022): Augmented successor of OK-VQA dataset, focused on common-sense knowledge and reasoning rather than purely factual knowledge, A-OKVQA offers approximately 24,903 questions across 23,692 images. Questions in this dataset demand common-sense reasoning about the scenes depicted in the images, moving beyond straightforward knowledge base queries. It also provides rationales for answers, aiming to be a significant testbed for the development of AI models that integrate visual and natural language reasoning.

Dataset	Images Count	Questions Count	Knowledge Type
KB-VQA	700	2,402	Fixed KB
FVQA	2,190	5,826	Fixed KB
OK-VQA	14,031	14,055	Factoid (Open-ended KB)
A-OKVQA	23,692	24,903	Factoid/Common-Sense

Table 3-3 KB-VQA datasets.

Upon the examination of a range of datasets available in the KB-VQA domain, the OK-VQA dataset (Marino et al., 2019) was selected for the research. This choice is particularly driven by the dataset's recognition and extensive use within the AI research community (Gui et al., 2022; Lymperaïou and Stamou, 2023; Shao et al., 2023; Reichman et al., 2023). The widespread adoption of OK-VQA as a benchmark provides a significant opportunity to position and evaluate the model within the context of current state-of-the-art systems. Additionally, the diversity and complexity of the OK-VQA dataset are provided as a comprehensive platform for testing the capabilities of the model in various scenarios. An emphasis on questions that require reasoning beyond the visual content is aligned with the research aim to develop AI systems capable of advanced, multi-faceted problem-solving. Furthermore, the dynamic nature of its knowledge base, reflecting the evolving and open-ended nature of real-world information, presents an ideal setting for testing the adaptability and learning capabilities of AI models.

### 3.2.2. OK-VQA Dataset Analysis

OK-VQA dataset (Marino et al., 2019) has been meticulously designed to furnish a collection of question-answer pairs, leveraging a selected subset of images from the COCO dataset (Lin et al., 2014).

The authors extracted a random sample of 14,031 images from the COCO dataset, upon which they employed a group of ten<sup>9</sup> individuals to generate a question for each image. The criterion for these questions was that they should necessitate external knowledge for accurate answering. Subsequently, the questions deemed to require no external knowledge were filtered out. Following this, a different set of 10 individuals were tasked with providing answers to the remaining questions. The final step involved the elimination of question-answer pairs that exhibited a bias towards certain answers. The resulted dataset has 14,055 questions about the 14,031 images with 24 images having multiple questions.

OK-VQA dataset boasts an extensive variety of scenes, extracted from the COCO image repository, which are categorically distributed into 10 principal categories as illustrated in [Figure 3-3](#). A distinctive feature of OK-VQA is its openness; it is not constrained to a 'closed' dataset, nor is it exclusively derived from a singular, specified source (Zakari et al., 2022). Instead, it epitomizes the concept of 'open'-domain knowledge, presenting a broader spectrum of information and applications. The dataset has an average length of 8.1 words per question and an average of 1.3 words per answer, some of the dataset main characteristics are summarized in [Table 3-4](#).

<sup>9</sup> Although the original paper mentions 5 individuals (Marino et al., 2019), the actual dataset has 10 answers for each question.

QUESTIONS DISTRIBUTION OVER KNOWLEDGE CATEGORIES

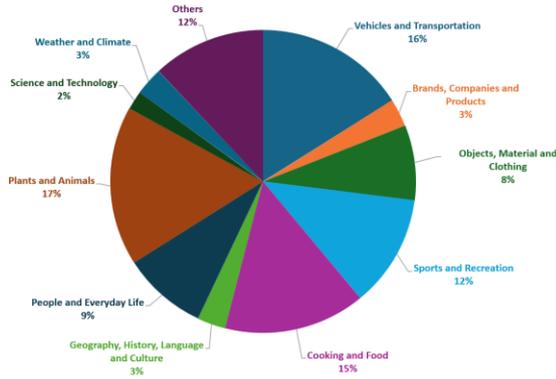


Figure 3-3 Questions distribution over knowledge categories.

OK-VQA Dataset Characteristics	
No. of Images	14,031
No. of Questions	14,055
No. of Questions Categories	10 + 1 (others)
No. of unique Entities	12,591
Ground Truth Answers per Question	10
Average Question Length	8.1 (words)
Average Answer Length	1.3 (words)
No. of Unique Questions	12,591
No. of Unique Answers	14,454
No. of Questions with Consensus Answers	1,701
No. of Unique Words in Questions	5,703
No. of Unique Words in Answers	11,125

Table 3-4 OK-VQA dataset characteristics.

The questions themselves are crafted to replicate real-life scenarios using natural language questioning keywords, Figure 3-4 shows the distribution of the questioning keywords used in the dataset questions.

A recent study by (Y. Chen et al., 2023) has found that approximately 70.8% of questions within the dataset can be answered by an average educated adult without the need for specific knowledge search. This suggests that the dataset was predominantly constructed with commonly known information, accessible to the general people.

During dataset investigation, further observations were made including:

- Spelling Mistakes:** (e.g. “sandwich” instead of sandwich and “moutains” instead of mountains).
- Grammar Mistakes:** (e.g. “What kind is train is that?” instead of “What kind of train is that?” or “In this image what believes soup is too hot too cold and just right?” Instead of “In this image, who believes the soup is too hot, too cold, or just right?”).
- Wrong Answers:** (e.g. for an image showing a traffic sign in Arabic, the question is “What language is this sign in?” and some of the answers were “Saudia Arabian” or “Japanese”)
- Nuanced Distinctions:** Some questions involved distinctions too subtle for individuals or even algorithms to discern, such as differentiating between types of oranges or two types of airplanes.
- Numerical Ranges:** some questions ask about numerical values that normally don’t have a fixed number (e.g. “How tall do these animals grow to be?” for an image showing a zebra).
- Ambiguous or Incorrect Questions:** Some questions are ambiguous or even incorrect (e.g. “What year will the vehicle? “).
- Contradicting Ground Truth Answers:** Many of the questions have ground truth answers that are contradicting with each other (e.g. “Electric stove” vs “Gas stove” or “Summer” vs “Winter”).

While it can be argued that the presence of contradicting ground truth or wrong answers may challenge the evaluation process, the remaining observations within the dataset do not undermine its validity. On the contrary, these elements—such as spelling mistakes, grammatical errors, nuanced distinctions, and variable numerical ranges—authentically reflect the complexity and variability encountered in real-world scenarios. Thus, despite some inaccuracies, this dataset remains a highly appropriate choice for simulating real-life environments, where information is often imperfect and demands critical interpretation and analysis. The inherent imperfections, rather than detracting from its value, enhance the dataset’s utility in developing more

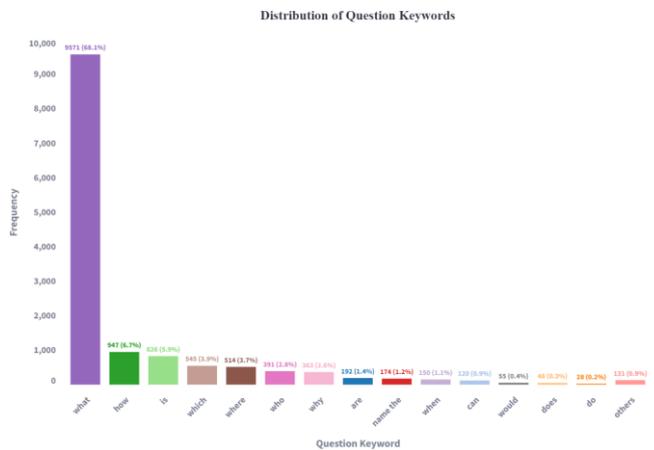


Figure 3-4 OK-VQA question keyword distribution.

resilient and adaptable artificial intelligence systems capable of navigating and interpreting the intricacies of human language and knowledge.

## 3.3. Implementation

### 3.3.1. Component Models

In the implementation of the object detection stage, the KB-VQA system explores different configurations for experimental purposes, one configuration utilizes the 'Yolov5-small' (Jocher, 2020) model from Ultralytics<sup>10</sup>, another experimental setup employs DETIC (Zhou et al., 2022) from Meta via HuggingFace Transformers library<sup>11</sup>, this setup corresponds to the 'Detic\_DeformDETR\_R50\_4x' checkpoint from the original repository<sup>12</sup>, for both configurations, the default settings were retained except for the 'Confidence Threshold', which was adjusted to 0.2 during both fine-tuning and testing stages. All images' captions were generated using InstructBLIP (Dai et al., 2023) via HuggingFace Transformers library<sup>13</sup>. The PT-LLM used was LLaMA-2 Chat (Touvron et al., 2023b) with two configurations, 'LLaMA-2 7B Chat' and 'LLaMA-2 13B Chat'. Detailed configurations for all the component models can be found in [Appendix B.2](#).

### 3.3.2. Fine-Tuning

A critical aspect of the KB-VQA system design involves two fundamental requirements: firstly, guiding the PT-LLM model in interpreting language-mediated visual contexts and leveraging its reasoning capabilities to extract the necessary implicit knowledge for answering queries; secondly, ensuring that the model's responses conform to the specific answer format required by the OK-VQA dataset, which is primarily limited to 1-2 words as demonstrated in [Table 3-4](#). Predominantly, existing research (Yang et al., 2022; Shao et al., 2023; Tan and Shen, 2023; Xenos et al., 2023) has adopted In-Context (n-shot) learning strategies, where a text and vision encoder modules of models like CLIP (Radford et al., 2021) or BLIP (J. Li et al., 2022) are employed to generate embeddings for a selected set of examples, at each inference stage, n-samples are selected from the complete set based on their cosine similarity to the current input's embedding, and used to exploit the PT-LLM n-shot learning capabilities. This approach, however, faces several challenges. First and foremost, the selection of relevant examples for each new query is a significant hurdle due to the finite number of prepared examples, which might not cover all possible question-image scenarios comprehensively. Additionally, the PT-LLM's context window imposes a significant restriction on the number of examples that can be effectively utilized at each inference, limiting the scope of contextual understanding. Thirdly, the computational demand of processing multiple examples for each inquiry can be computation expensive, thus affecting the system's efficiency and scalability.

In contrast, this project implements a custom instruction-based fine-tuning approach for the PT-LLM. By integrating the custom special tokens discussed in [Section 3.1.3](#) into the model's vocabulary and customizing the training data based on a structure that guides the model on how to interpret these special tokens, the modality gap effectively shrinks and the model becomes more visually aware of the scene. The OK-VQA dataset was re-engineered to incorporate the objects and captions and re-structure all entries to follow this structured format as illustrated in [Figure 3-5](#).

---

<sup>10</sup> <https://github.com/ultralytics/ultralytics>

<sup>11</sup> <https://huggingface.co/facebook/deformable-detr-detic>

<sup>12</sup> <https://github.com/facebookresearch/Detic?tab=readme-ov-file>

<sup>13</sup> [https://huggingface.co/docs/transformers/main/en/model\\_doc/instructblip#transformers.InstructBlipForConditionalGeneration](https://huggingface.co/docs/transformers/main/en/model_doc/instructblip#transformers.InstructBlipForConditionalGeneration)

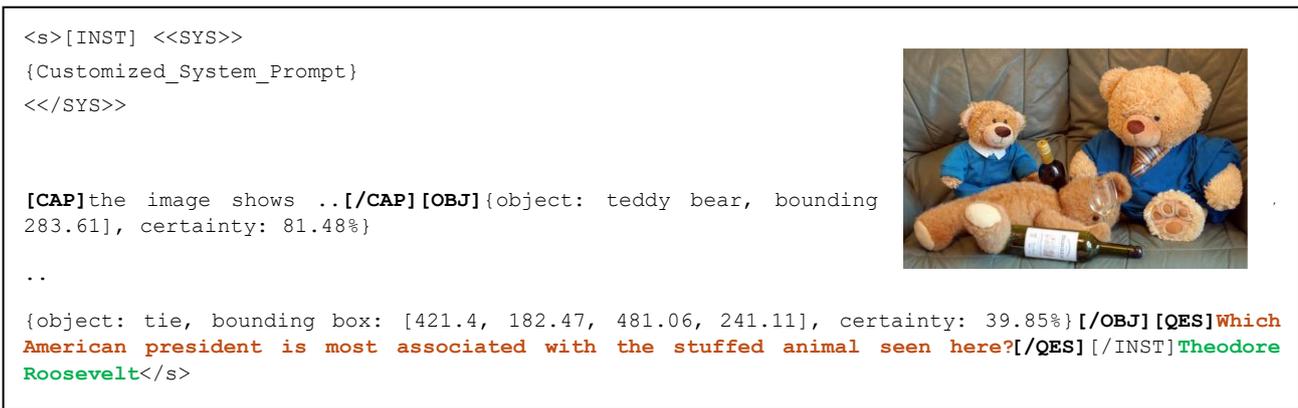


Figure 3-5 Fine-tuning data structure.

This method addresses the limitations of n-shot learning and not only meets the need for sophisticated reasoning over visual and textual data but also ensures that the outputs strictly adhere to the dataset's concise answer format without excessive computational overhead.

The fine-tuning process began by structuring 8,135 samples from the dataset according to the format depicted in Figure 3-5, where the most popular answer among the ground truth answers was used together with the objects list detected using YOLOv5. To enable the fine-tuning on a single GPU with the lowest GPU RAM requirements, all samples exceeding a token count of 1,024 were excluded, resulting in 7,403 viable samples. Subsequently, a Parameter-Efficient Fine-tuning (PEFT)<sup>14</sup> approach was employed using QLORA (Quantized Low Rank Adapter) (Dettmers et al., 2023). This approach combines quantization—reducing numerical precision from 32-bit to 16-bit or 4-bit floating points—with LORA (Low Rank Adapter) (Hu et al., 2021), which introduces a trainable adapter layer over the LLM's existing layers. Critically, this method does not require training the entire model's weights but focuses solely on the adapter<sup>15</sup>. This quantized adapter has been shown to be efficient while having a minimal impact on model performance (Zhao et al., 2023; Naveed et al., 2023). The model was fine-tuned using 4-bit quantization for one epoch and a decaying learning rate of  $2 \times 10^{-4}$ , with 201.9 million and 129 million trainable parameters for LLaMA-2 13B and LLaMA-2 7B, respectively. Fine-tuning configuration details and results can be found in Appendix B.3.

### 3.3.3. Hardware & Environment

All project stages were implemented on Google Colab Pro+. A single T4 GPU – 15 GB GPU RAM was used during the whole research except for the fine-tuning stage where a single A100 GPU – 40GB GPU RAM was used.

GPU requirements were calculated using the below formula (Stoelinga, 2023):

$M = \frac{(P * 4B)}{(32/Q)} * 1.2$	M: Required GPU in GB. 4B: 4Bytes (32 bits). 1.2: Additional Overhead.	P: Model Size. Q: Quantization (e.g. 8-bit, 4-bit).
-------------------------------------	--	--

Equation 3-1 GPU requirements calculation.

<sup>14</sup> <https://github.com/huggingface/peft>

<sup>15</sup> Fine-tuning GPT-3 175B using LORA reduces the trainable parameters by up to 10,000 times (Hu et al., 2021).

### 3.3.4. Interactive Model Access on Huggingface Space

To improve the accessibility and demonstrability of the Knowledge-Based Visual Question Answering (KB-VQA) model, a detailed demonstration has been set up on Hugging Face Spaces. It can be accessed through the link: <https://huggingface.co/spaces/m7mdal7aj/KB-VQA>

This platform offers an intuitive interface for users to interact with the KB-VQA model's features, including the fine-tuned models available for immediate use and assessment. The complete code for the project is also accessible in this space, allowing for easy review and customization. The demonstration not only presents the project in its entirety but also aids in understanding the foundational processes. It provides in-depth views of the training and validation dataset, highlighting the variety and characteristics of the data. Additionally, the model's performance is thoroughly documented through evaluation results, providing concrete proof of its effectiveness across different measures. The model's testing stage is comprehensively detailed, giving users the opportunity to test the model's real-time response to a range of visual questions. This extensive demonstration is a crucial tool for both scholars and industry professionals, enhancing openness and supporting ongoing research and innovation efforts.

## Chapter 4

# 4. Evaluation and Results

Evaluating the KB-VQA model on the OK-VQA dataset, which focuses on open-ended questions, poses unique challenges compared to other datasets that use multiple-choice formats (Lu et al., 2022). Traditional metrics fall short in capturing the complexities of open-ended responses, which require not only correctness but also contextual relevance and deep semantic understanding. Thus, a more sophisticated evaluation framework is essential to accurately assess the model's performance in this nuanced domain.

### 4.1. Evaluation Metrics

In the domain of Knowledge-Based Visual Question Answering (KB-VQA), the evaluation of model efficacy transcends the realm of straightforward quantitative analysis, necessitating a multifaceted approach to accurately gauge performance. The inherent complexity of KB-VQA models, which integrate intricate aspects of both visual perception and knowledge-based reasoning, poses significant challenges in their assessment. Unlike conventional tasks where a singular, definitive answer is often expected, KBVQA scenarios are characterized by a diverse array of potential correct answers, each varying in their semantic and contextual appropriateness. This multiplicity of valid responses, coupled with the subjective nature of certain queries, renders traditional evaluation metrics such as binary accuracy insufficient. Consequently, the evaluation process must be meticulously designed to encompass not only the correctness of the answers in a literal sense but also their relevance and alignment with human-like understanding and reasoning. This necessitates the adoption of advanced, nuanced metrics that can effectively capture the depth and breadth of the model's cognitive and interpretative capabilities, thereby providing a more holistic and representative assessment of its performance in KBVQA tasks. Below are several existing metrics that are commonly employed in the evaluation of KB-VQA system:

- 1. Simple Accuracy:** The Simple Accuracy metric, commonly used in various ML systems including question-answering model, finds limited application in KB-VQA, it can be applicable when when dealing with datasets like that feature questions with multiple-choice answers such as A-OKVQA (Schwenk et al., 2022) or ScienceQA (Lu et al., 2022).

$$\text{Simple Accuracy} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Answers}} \times 100\% \quad \text{Equation 4-1}$$

However, this metric's effectiveness diminishes in the context of open-ended questions, typical of datasets like OK-VQA. Such questions, reflecting real-world complexities, present a wide array of potential correct answers, challenging the Simple Accuracy metric's capacity to capture the nuances of human cognition. This limitation underscores the necessity for more sophisticated, contextually adaptive evaluation methods in KBVQA, capable of accommodating the diverse range of plausible answers inherent in open-ended

scenarios, thereby providing a more accurate reflection of a model's ability to mimic human-like understanding and reasoning.

**2. VQA Score (Soft Accuracy)** (Antol et al., 2015): VQA Score, also known as the Consensus Metric (Zakari et al., 2022) stands as the most commonly and widely used metric for evaluating the majority of KB-VQA models (Wu et al., 2021; Lin et al., 2022; Gui et al., 2022; Yang et al., 2022; Z. Chen et al., 2023; Shao et al., 2023).

This metric, predicated on a consensus-based approach, deems an answer to be 100% accurate if it aligns with at least three out of the ten provided ground truth answers for each question. This approach is particularly adept at accommodating the inherent variability and subjectivity of human responses in KB-VQA tasks.

VQA score can be broken down into two formulas. The first formula calculates the accuracy for each individual evaluation sample, while the second formula aggregates these accuracies to provide an overall Soft Accuracy score. This scoring criteria gives the model a partial credit, even if it generates an answer that is less common among ground-truth answers.

$$\text{Soft Accuracy}_{\text{sample}} = \min \left( \frac{\text{Number of Matching Ground Truth Answers}}{3}, 1 \right) \quad \text{Equation 4-4}$$

$$\text{Soft Accuracy} = \frac{\sum_{i=1}^N \text{Accuracy}_{\text{sample}_i}}{N} \quad \text{Equation 4-3}$$

Where  $N$  is the total number of the evaluation samples and  $i$  is the  $i^{\text{th}}$  sample.

The utilization of this metric is predominantly chosen for its simplicity (de Faria et al., 2023) and the ease of access to 10 ground-truth answers in certain datasets, such as OK-VQA (Marino et al., 2019). Nevertheless, this approach is not without limitations. The collection of 10 ground-truth answers for each question poses a significant expense. Moreover, achieving a consensus among 3 annotators on most questions is a challenging endeavour. This difficulty escalates markedly in obtaining unanimous responses from 3 annotators for 'why' questions or subjective inquiries necessitating a numerical response. Such constraints inherently cap the maximum evaluation score attainable by a model.

Moreover, the employment of this metric for model training, as opposed to mere evaluation, presents another caveat. It may inadvertently prompt models to generate predictable, commonplace responses that align more closely with the ground truth. This tendency potentially occurs at the expense of more innovative or insightful answers, which, although equally valid, may be less prevalent. The implications of this on the development of models warrant careful consideration, as it could lead to a propensity for safe, but less informative, predictions.

**3. Exact Match (EM)**: The Exact Match metric, as defined by (Gao et al., 2022b), is a specific evaluation method used in Visual Question Answering (VQA) that measures the accuracy of a VQA system based on the exactness of its predictions compared to a set of annotated answers. Unlike a strict one-to-one comparison with a single ground truth, the EM metric considers a prediction correct if it exactly matches any one of the 10 annotated answers provided for each question. This approach acknowledges the possibility of multiple valid answers for a single question in VQA tasks.

The EM metric is calculated as the percentage of questions for which the predicted answer is an exact match to at least one of the 10 provided annotated answers. This method treats all annotated answers as equally valid ground truths, thereby offering a more flexible and realistic assessment of a model's performance in scenarios where a range of correct answers is possible. The EM metric was utilized in evaluating KB-VQA models such as TRiG (Gao et al., 2022b).

While this metric is considerably less stringent than the VQA score, it still focuses on the exact match of the answer, potentially missing the semantic accuracy where different phrasing or synonyms could convey the correct answer. This might lead to undervaluing models that provide semantically correct but differently

phrased responses. Additionally, this metric may lead to an over-simplification of the evaluation process. In real-world scenarios, certain answers, while technically correct, might be less informative or less useful than others. Furthermore, this metric doesn't differentiate between frequent and less frequent answers, treating them with equal weight.

**4. Answer Frequency Rank (Marino et al., 2021):** This metric is employed to assess the alignment of model predictions with commonly occurring answers in a reference dataset. This metric ranks potential answers based on their frequency of occurrence, with the premise that more frequent answers are more likely to be correct in certain contexts. This metric was utilized in evaluating KB-VQA models like KRISP (Marino et al., 2021).

The implementation of this metric involves two key steps. First, each answer in the reference set is assigned a Frequency-based Rank,  $FR$ , where the most frequent answer is given the highest rank. Then, a Rank Score

$$RS(r) = N - r + 1 \quad \text{Equation 4-5}$$

( $RS$ ) is allocated to each rank. For a scenario with  $N$  distinct ranks, the score for rank  $r$  is defined as:

The overall performance score  $S$  for the VQA model is calculated as the average of these Rank Scores across all questions, using the formula:

$$S = \frac{1}{Q} \sum_{i=1}^Q RS(r_i) \quad \text{Equation 4-6}$$

where  $Q$  is the total number of evaluation questions, and  $r_i$  is the rank of the model's predicted answer for the  $i^{th}$  question.

The Answer Frequency Rank metric, while providing a quantitative measure of a model's ability to predict commonly occurring human responses, has its share of limitations. One of the issues is its sensitivity to the distribution of answers within the dataset, making it highly sensitive to any bias in the dataset (de Faria et al., 2023). Consequently, models might excel according to this metric but struggle in handling novel or infrequent responses. Moreover, the metric lacks nuance in differentiating between the semantic diversity of answers, treating all frequent responses equally, regardless of their informativeness. This approach overlooks the distinction between trivial and substantial answers, leading to an incomplete evaluation of a model's capabilities. Another significant limitation is the metric's penalization of models for predicting rare, yet contextually pertinent, answers, thereby ignoring their relevance in specific scenarios. Additionally, the metric's assumption that frequently occurring answers are more likely to be correct does not consider the variability of context, where an answer valid in one situation might be inapplicable in another.

## 4.2. Evaluation Process

In the evaluation phase of the Knowledge-Based Visual Question Answering (KB-VQA) model, the assessment was structured to encompass both syntactic and semantic evaluations, employing VQA Score (Antol et al., 2015) and Exact Match (Gao et al., 2022b) as the primary scoring criteria. A representative sample of 1,000 image-question pairs, evenly distributed across all question categories and excluded from finetuning data, was selected for this comprehensive analysis.

### 4.2.1. Syntactic Evaluation

The syntactic evaluation, aimed at aligning with state-of-the-art KB-VQA models, focuses on the literal alignment of words and phrases. To ensure fair comparison, a meticulous process emphasizing syntactic accuracy was adopted, which is crucial for facilitating rigorous comparisons between the evaluated model's outputs and the

ground truth answers. This process adheres to the common practice in KB-VQA models evaluation models (Wu et al., 2021; Lin et al., 2022; Gui et al., 2022; Yang et al., 2022; Z. Chen et al., 2023; Shao et al., 2023), prioritizing syntactic precision to maintain the integrity and uniformity of the assessment.

The process involved applying the following steps to the ground truth and KB-VQA model answers:

1. **Text Normalization:** All text is standardized to a uniform format, primarily by converting it to lowercase.  
Tool Used: Python's string methods.  
Example: Equating "Earth" with "earth".
2. **Stemming:** Words are converted to their base form, which standardizes various forms of the words.  
Example: "biking" and "bikes" are stemmed to "bike".  
Tool Used: *NLTK's PorterStemmer*.
3. **Compound Words and Hyphenated Terms:** Compound words and hyphenated terms are reconciliated.  
Example: Matching "racecar" with "race car" and "teddy bear" with "teddy-bear".  
Tool Used: *FuzzyWuzzy* library followed by manual review.
4. **Spelling Mistakes:** Minor spelling mistakes are considered a match.  
Example: "bicylc" with "bicycle" and "sunlihjy" with "sunlight".  
Tool Used: *FuzzyWuzzy* library followed by manual review.
5. **Simple Variations:** Simple variation is considered a match.  
Example: "grapes and bananas" vs "bananas and grapes".  
Tool Used: *FuzzyWuzzy* library followed by manual review.

This approach strategically excluded semantic considerations to benchmark against other state-of-the-art KB-VQA models. Synonyms and semantic equivalents, like 'automobile' versus 'car' or "texting" versus "chatting", were not recognized as matches to maintain a focus on syntactic precision.

During the syntactic-focused evaluation across various experimental setups, after applying the outlined preprocessing steps to both the model and ground-truth answers, 86 samples were identified as potential matches based on a *FuzzyWuzzy ratio* > 80%. Each of these 86 samples was subjected to a thorough manual review to validate the accuracy of the match. Through this meticulous process, approximately 30% of these initially identified matches were subsequently reclassified as mismatches. Notable examples of such mismatches include matching "15 years" with "14 years" or "50 feet" with "500 feet".

#### 4.2.2. Semantic Evaluation

Semantic evaluation recognizes the importance of context and meaning, going beyond the literal alignment of words. This approach is particularly effective in assessing the model's ability to generate accurate answers that are not just syntactically correct but also semantically relevant.

In addition to the syntactic-focused evaluation, a semantic evaluation was conducted using the same scoring criteria – VQA Score (Antol et al., 2015) and Exact Match (Gao et al., 2022b). However, this approach differed significantly in its consideration of both syntactic and semantic matching, inspired by (Mañas, Krojer and Agrawal, 2024) the advanced capabilities of Large Language Models (LLMs) were leveraged, specifically GPT-4 (OpenAI, 2023).

The necessity for this approach becomes evident when considering the diversity in human responses: In our dataset, 1,701 questions showed complete agreement among annotators, suggesting syntactic consistency. However, there were 12,354 questions with partial or no agreement, highlighting the limitations of a purely syntactic analysis. This disparity underscores the need for a semantico-syntactic<sup>16</sup> evaluation that not only assesses structural accuracy but also contextual and semantic appropriateness.

By employing LLMs, the evaluation process taps into a deeper understanding of language, considering nuances, context, and the intended meaning behind words and phrases. This methodology aligns with the ultimate goal

<sup>16</sup> Semantico-syntactic evaluation refers to an approach that combines both semantic and syntactic analysis in the assessment of language or language-based models.

of developing a KB-VQA model that delivers correct answers, irrespective of the specific syntax used. It acknowledges that in real-world scenarios, the effectiveness of a VQA model is determined not only by its adherence to syntactic accuracy but also by its capacity to interpret and respond to the semantic content of queries.

Leveraging LLMs for evaluating the KB-VQA is beneficial for many reasons including:

- 1. Contextual Understanding:** LLMs like GPT-4 offer a nuanced understanding of context. Their advanced algorithms enable them to interpret questions and answers in a human-like manner, leading to more accurate and realistic evaluations of the model's performance.
- 2. Flexibility in Language Use:** Users often phrase similar queries differently. LLMs' ability to process and understand this linguistic variability ensures that the evaluated model can accurately respond to a wide range of query formulations.
- 3. Comprehensive Model Assessment:** LLMs facilitate an evaluation that encompasses both syntactic and semantic considerations. This results in a holistic view of the model, assessing not only the correctness of the information provided but also its relevance and appropriateness.
- 4. Alignment with Real-World Usage:** Modern users expect conversational AI to grasp queries beyond just the literal word matching. LLMs enable this understanding, ensuring that the model is tested against standards that reflect actual use cases.

Nevertheless, this approach is not without limitations:

- 1. Cost:** Utilizing models like GPT-4 for extensive evaluation can be costly, GPT-4 API costs 90\$ for 1 million tokens input and 1 million tokens output (OpenAI, 2024a).
- 2. Potential for Hallucination:** Even well-trained models like GPT-4 can "hallucinate," i.e., generate plausible but incorrect or nonsensical information. This is particularly true for ambiguous questions or answers where the model might generate confident responses that are factually incorrect or irrelevant (Mañas, Krojer and Agrawal, 2024).
- 3. Handling Ambiguity:** Ambiguous questions or answers pose a significant challenge. GPT-4, while adept at processing language, might still struggle to derive clear, accurate conclusions from ambiguous or poorly structured inputs. This is a crucial consideration for KB-VQA models where the clarity and specificity of responses are essential.
- 4. Probabilistic Nature:** Being based on probabilistic models, LLMs like GPT-4 might produce different outputs for similar or identical inputs due to the inherent randomness in their response generation process. In an effort to control this as much as possible, OpenAI recommends using a fixed temperature in addition to the Beta feature of adding a seed for the model to try – but does not guarantee – to sample deterministically (OpenAI, 2024b)

Human evaluation remains the gold standard for its nuanced understanding; however, it falls short in practicality and scalability for large-scale datasets. In contrast, while PT-LLMs like GPT-4 have their limitations, their benefits — in terms of efficiency and breadth of language comprehension — significantly outweigh these limitations, making them a highly practical alternative for evaluating KB-VQA models. It can also be argued that standard evaluation metrics, like the VQA score (Antol et al., 2015) which primarily focuses on syntactic equivalence, may be more appropriate for evaluating vanilla VQA models (Antol et al., 2015). These models are designed to address VQA tasks where the required information is solely derived from the image content itself. However, this approach becomes less efficient when applied to KB-VQA models that necessitate external knowledge beyond the image content. This requirement introduces a level of subjectivity to the knowledge, which the standard metrics are not equipped to assess accurately.

Semantic evaluation was conducted using GPT-4 API employing the following template and configurations:

<b>Seed</b>	123
<b>Temperature</b>	0.1
<b>Template</b>	<pre> {"role": 'system', 'content': "You are an AI trained to evaluate the equivalence of AI-generated answers to a set of ground truth answers for a given question. Upon reviewing a model's answer, determine if it matches the ground truths. Use the following rating system: 1 if you find that the model answer matches more than 25% of the ground truth answers, 2 if you find that the model answer matches only less than 25% of the ground truth answers, and 3 if the model answer is incorrect. Respond in the format below for easy parsing: Rating: {1/2/3}" {'role': 'user', 'content': "Question: What breed of dog is seen in this picture? Ground Truth: ['bulldog', 'bulldog', 'bulldog', 'bulldog', 'boxer', 'boxer', 'boxer', 'boxer', 'mongrel', 'mongrel'] Model's Response: boxer"} </pre>

Table 4-1 GPT-4 API settings for semantic evaluation.

### 4.3. Main Results

Table 4-2 summarizes the syntactic (employing string matching) and semantic (utilizing GPT-4) evaluations of the KB-VQA model’s performance, leveraging both metrics the VQA score and Exact Match (EM) score on the OK-VQA dataset, as delineated in Section 4.2. The evaluations encompass four distinct configurations, integrating two sizes of the LLaMA-2 model (7B and 13B) with two object detection models (DETIc and YOLOv5). The configuration combining LLaMA-2 13B with DETIc demonstrates superior performance across both syntactic and semantic evaluations, as detailed in the last row of the table.

Model	Syntactic Evaluation (%)		Semantic Evaluation (%) (GPT-4)			
	VQA Score	EM Score	VQA Scores	$\Delta_V$	EM Score	$\Delta_E$
<b>ConceptBert</b> (Gardères et al.,	33.66	-	-	-	-	-
<b>MAVEx</b> (Wu et al., 2021)	39.4	-	-	-	-	-
<b>KRISP</b> (Marino et al., 2021)	38.9	-	-	-	-	-
<b>PICa</b> (Yang et al., 2022)	48	-	-	-	-	-
<b>KAT</b> (Gui et al., 2022)	54.41	-	-	-	-	-
<b>REVIVE</b> (Lin et al., 2022)	58	-	-	-	-	-
<b>MuKEA</b> (Ding et al., 2022)	42.59	-	-	-	-	-
<b>TRiG</b> (Gao et al., 2022b)	50.5	54.73	-	-	-	-
<b>IPVR</b> (Z. Chen et al., 2023)	44.62	-	-	-	-	-
<b>PROPHET</b> (Shao et al., 2023)	61.1	-	-	-	-	-
<b>LAMOC</b> (Du et al., 2023)	40.31	-	-	-	-	-
<b>TwO</b> (Si et al., 2023)	58.72	61.32	-	-	-	-
<b>Q&amp;APrompts</b> (Wang and Ge,	<b>64.3</b>	-	-	-	-	-
<b>GeReA</b> (Ma et al., 2024)	<b>66.5</b>	-	-	-	-	-
<b>KB-VQA (This Project)</b>						
<b>7B (Caption + YOLOv5)</b>	57.19	61.08	65.99	8.78	67.86	6.78
<b>7B (Caption + DETIc)</b>	62.51	67.07	70.19	7.68	71.96	4.89
<b>13B (Caption + YOLOv5)</b>	60.15	64.77	68.86	8.71	70.66	5.89
<b>13B (Caption + DETIc)</b>	<b>63.57</b>	<b>68.36</b>	<b>71.09</b>	7.52	<b>72.55</b>	4.19

Table 4-2 Main results of the KB-VQA system compared to existing methods.

$\Delta_V, \Delta_E$ : Difference between syntactic and semantic evaluation of VQA and EM scores, respectively.

Significantly, both EM and VQA scores show notable improvements in the semantic evaluation compared to the syntactic assessment. For instance, the optimal model configuration, 13B (Caption + DETIC), registers a +7.52% increase in the VQA score and a +4.19% rise in the EM score, indicating superior prediction of semantically correct answers, which aligns closely with real-world application needs.

Utilizing DETIC as the object detector, in contrast to YOLOv5, yields significant improvements in the syntactic evaluation, with increases of +5.3% and +3.42% in the VQA score for the 7B and 13B model sizes, respectively, and enhancements of +5.99% and +3.59% in the EM score. Comparable gains are also noted in the semantic evaluation. These results substantiate the design choice to integrate an object detector capable of detecting a broader spectrum of object classes (DETIC versus YOLOv5). The performance gains from DETIC over YOLOv5 are more pronounced with the smaller LLaMA-2 size (7B versus 13B), suggesting that DETIC’s enriched visual context compensates for the inherent limitations of the smaller model size.

Across different LLaMA-2 model sizes, regardless of the object detection model used, evaluation criteria, or metric, it is evident from the results that larger models yield better results. For example, the 13B (Caption + YOLOv5) configuration achieves a 64.77% (+3.69%) EM score, outperforming the 61.08% scored by the 7B (Caption + YOLOv5) configuration. This trend is consistent across all model configurations and evaluation metrics. Larger PT-LLMs demonstrate stronger language modeling capabilities, facilitating a more profound comprehension of the query, a wealthier repository of implicit knowledge, and more effective reasoning across the entire input (query, caption, and objects). These findings align with previous research (Guo et al., 2023; Xenos et al., 2023), which reported similar benefits when using larger PT-LLMs in KB-VQA systems, and call for further experimentation with the largest available LLaMA-2 size, the 70B.

#### 4.4. Comparative Results

Although certain studies (Lin et al., 2022; Ma et al., 2024) have implied that semantic evaluation of models might lead to more accurate results, the prevailing norm continues to be the standard syntactic evaluation. Consequently, these studies have not reported their findings using semantic metrics. Therefore, the comparative analysis remains confined to syntactic evaluations for both VQA and EM scores.

In terms of the EM score, the model demonstrates superior performance, surpassing those models previously benchmarked using this metric by achieving 68.36%, compared to 61.32% for TwO (+7.04%) and 54.73% for Trig (+13.63%). In terms of the VQA score, the model displays highly competitive results, surpassing nearly all existing models except for GeReA, which leads by 2.93%, and Q&APrompts, by 0.73%. The superior results of GeReA can be attributed to the richer visual context resulting from the employment of two large models (InstructBLIP + LLaVA-1.5), where each generates multiple captions for each key region in the image, providing a deeper and more nuanced understanding. However, this approach may limit efficiency during inference due to increased computational demands. The assertion that the superior performance of GeReA is attributed to the use of dual models can be substantiated by examining the results of GeReA when employing a single model. GeReA developers report their results when using a single captioning model to be 62.1% for InstructBLIP (i.e. inferior to the developed model in this research) and 63.6% for LLaVA-1.5 (approximately similar to this research model results). Conversely, the negligible difference of 0.73% between this research model and Q&APrompts indicates a comparable level of performance, likely within the margin of experimental error or minor inherent testing variances.

While the KB-VQA system developed in this research is shown to demonstrate superior performance on the EM score relative to models that have reported results using this metric, it must be acknowledged that not all models have been evaluated using the EM score. As the VQA score is the uniform metric across models, the EM score provides additional insights but cannot be utilized to definitively benchmark the system against all existing models.

### 4.5. Qualitative Analysis

Table 4-3 showcases a selection of examples from the validation data, depicting both successful and unsuccessful model predictions. The model utilizes visual context for comprehending questions and deriving answers, drawing from either generated image captions or detected objects within bounding boxes. Notably, the initial three rows demonstrate successful model predictions. This success is attributed to the model’s engagement in reasoning processes that harness key terms from the visual context, supplemented by the retrieval of pertinent knowledge. For instance, in the first row, the model adeptly identifies an animal as a 'cat' from visual cues present in both the caption and object detection. Utilizing its implicit knowledge base, it correctly deduces that the typical sound made by a content cat is purring, showcasing its ability to synthesize information from visual inputs and inherent knowledge to form accurate conclusions.

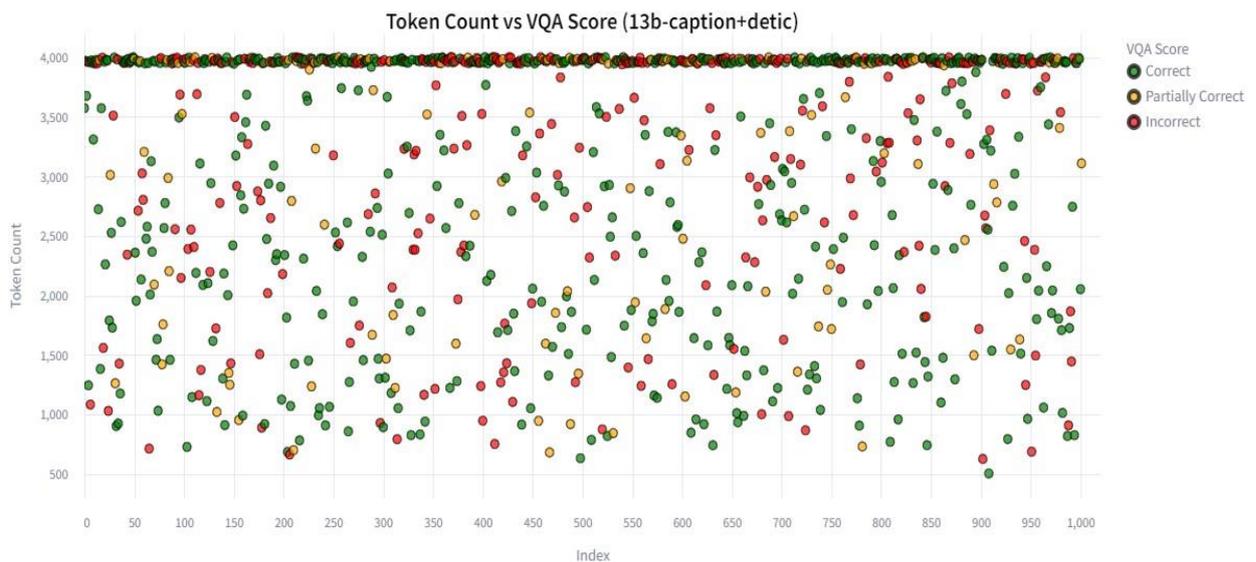
	<p><b>Q:</b> What sort of sound does this animal make when it's contented as it appears to be now?</p> <p><b>GTs:</b> ['purr', 'purr', 'purr', 'purr', 'purr', 'purr', 'purrs', 'purrs', 'meaow', 'meaow']</p> <p><b>Caption:</b> a cat is lying in a suitcase, which is open and partially filled with clothes ...</p> <p><b>Objects:</b> cat, pillow, crib, pet, cushion, kitten, ...</p> <p><b>Model Answer:</b> purr</p>	<p>VQA Score: 100%</p> <p>EM Score: 100%</p> <p>VQA Score (GPT-4): 100%</p> <p>EM Score (GPT-4): 100%</p>
	<p><b>Q:</b> What sort of hat is this man wearing?</p> <p><b>GTs:</b> ['cowboy', 'cowboy', 'cowboy', 'cowboy', 'cowboy', 'cowboy', 'cowboy', 'cowboy', 'cowboy', 'cowboy']</p> <p><b>Caption:</b> a man is standing near a body of water, wearing a hat and ...</p> <p><b>Objects:</b> hat, cellular_telephone, cowboy_hat, boat, ...</p> <p><b>Model Answer:</b> cowboy</p>	<p>VQA Score: 100%</p> <p>EM Score: 100%</p> <p>VQA Score (GPT-4): 100%</p> <p>EM Score (GPT-4): 100%</p>
	<p><b>Q:</b> In which US states are these fruits commonly grown?</p> <p><b>GTs:</b> ['florida', 'florida', 'florida', 'florida', 'florida', 'florida', 'florida', 'florida', 'florida and california', 'florida and california']</p> <p><b>Caption:</b> there are several oranges hanging ...</p> <p><b>Objects:</b> orange_(fruit), mandarin_orange, ...</p> <p><b>Model Answer:</b> florida</p>	<p>VQA Score: 100%</p> <p>EM Score: 100%</p> <p>VQA Score (GPT-4): 100%</p> <p>EM Score (GPT-4): 100%</p>
	<p><b>Q:</b> How many grams of sugar are in this item?</p> <p><b>GTs:</b> ['20', '20', '20', '25', '25', '4', '4', '6', '6']</p> <p><b>Caption:</b> a bagel is sitting on a white plate ...</p> <p><b>Objects:</b> bagel, plate, ...</p> <p><b>Model Answer:</b> 10</p>	<p>VQA Score: 0%</p> <p>EM Score: 0%</p> <p>VQA Score (GPT-4): 0%</p> <p>EM Score (GPT-4): 0%</p>
	<p><b>Q:</b> What kind of animal is this?</p> <p><b>GTs:</b> ['horse', 'horse', 'horse', 'horse', 'donkey', 'donkey', 'it is lama', 'it is lama', 'pony', 'pony']</p> <p><b>Caption:</b> a horse is standing in a grassy field ...</p> <p><b>Objects:</b> pony, pole, horse, ...</p> <p><b>Model Answer:</b> pony</p>	<p>VQA Score: 67%</p> <p>EM Score: 100%</p> <p>VQA Score (GPT-4): 100%</p> <p>EM Score (GPT-4): 100%</p>
	<p><b>Q:</b> What is a good side dish for this meal??</p> <p><b>GTs:</b> ['fries', 'fries', 'fries', 'fries', 'chips', 'chips', 'baked beans', 'baked beans', 'coleslaw', 'coleslaw']</p> <p><b>Caption:</b> a close-up shot of a sandwich is displayed on a piece of aluminium foil ...</p> <p><b>Objects:</b> sandwich, hamburger, beef_(food), ...</p> <p><b>Model Answer:</b> potatoes</p>	<p>VQA Score: 0%</p> <p>EM Score: 0%</p> <p>VQA Score (GPT-4): 67%</p> <p>EM Score (GPT-4): 100%</p>
	<p><b>Q:</b> What style of motorcycle is this called?</p> <p><b>GTs:</b> ['trike', 'trike', 'trike', 'trike', 'tricycle', 'tricycle', 'tricycle', 'tricycle', '3 wheeler', '3 wheeler']</p> <p><b>Caption:</b> a motorcycle parked on the side of the road ...</p> <p><b>Objects:</b> motorcycle, person, car ...</p> <p><b>Model Answer:</b> harley</p>	<p>VQA Score: 0%</p> <p>EM Score: 0%</p> <p>VQA Score (GPT-4): 0%</p> <p>EM Score (GPT-4): 0%</p>

Table 4-3 Qualitative visualization of the results. Predicted answers by the model configuration 13B (Caption + DETIC).

Further examination was conducted on 100 samples from the evaluation data where the model registered zero or partial VQA scores, to discern the underlying causes of these failures. The observations are as follows:

1. **Specialized Knowledge Requirement:** The model faces challenges with queries necessitating specialized knowledge not commonly known. An example is illustrated in row 4, where determining the sugar content in a bagel likely requires specialized nutritional knowledge. This highlights the necessity for targeted training, particularly in specialized domains, ensuring the model acquires the requisite depth of knowledge.
2. **Semantic Equivalence:** Reasonably correct responses from the model are frequently penalized due to the ground truth labels demanding specific terminology. As illustrated in row 5, responses that are semantically correct—even arguably more precise—but phrased differently are not fully acknowledged, impacting the model’s performance metrics.
3. **Ambiguous Questions with Subjective Answers:** Queries that are inherently ambiguous or yield multiple valid interpretations present significant challenges, as they complicate the model’s ability to select the most appropriate response. The query in row 6, which could elicit various subjective yet correct responses, exemplifies this. The model scored 0% in both VQA and EM metrics when evaluated syntactically and received partial credit when assessed semantically, with GPT-4 associating the term 'potatoes' with 'chips'.
4. **Loss of Visual Information During Image to Language Transformation:** A notable limitation is observed in the model’s capability to capture essential visual cues during the transformation of image content into language. This is particularly critical in images containing brand names, unclear text, or when distinctions in subtle nuances are necessary, as seen in the sample from row 7. This issue calls for further research aimed at enhancing the model's representation of the visual context.
5. **Bias, Inconsistency, or Subjectivity in Ground Truth Answers:** Despite the model providing logically or factually correct responses, discrepancies, subjectivity, and bias within the ground truth answers—as discussed in [Section 3.2.2](#)—result in partial or zero scoring. This misalignment reinforces the necessity to refine validation measures to accommodate semantic equivalences, as addressed in [Section 4.2.2](#).

To investigate the impact of the sequence length, the instances of successful and unsuccessful predictions by the model were correlated with the number of tokens for each sample inclusive of the question, caption, objects and bounding boxes, and there was no notable impact of the token count on the model performance as shown in [Figure 4-1](#).



*Figure 4-1 Token Counts vs VQA Score.*  
 VQA score displayed for the best performing model configuration: 13B (Caption + DETIC).  
 The high density of samples containing approximately 4,000 tokens results from trimming samples that exceeded the set threshold of 4,000 tokens.

## 4.6. Ablation Study

To systematically unpack the performance gains and understand the impact of different components within the model, and to expand on the reported performance for the four main model configurations discussed in [Section 4.3](#), an ablation study was conducted comparing various model architectures. Results of this study are detailed in [Table 4-4](#), which presents the performance metrics from these ablation experiments using syntactic evaluation<sup>17</sup> for both VQA and EM scores.

LLaMA-2 Model		Caption	Object Detector		Syntactic Evaluation (%)	
7B	13B		YOLOv5	DETIC	VQA Score	Exact Match
✓		✓			56.72	61.38
✓			✓		38.29	41.52
✓				✓	40.19	44.01
✓		✓	✓		57.19	61.08
✓		✓		✓	62.51	67.07
	✓	✓			57.49	62.28
	✓		✓		41.42	44.91
	✓			✓	42.81	46.81
	✓	✓	✓		60.15	64.77
	✓	✓		✓	63.57	68.36

Table 4-4 Ablation experiments for KB-VQA model components.

The experimental configurations included limiting the visual context to either captions alone, without objects, or to objects detected alone, without captions, and examining the use of specific detection models (either YOLOv5 or DETIC) both in isolation and in combination with captions. These experiments were executed utilizing both the 13B LLaMA-2 and 7B LLaMA-2 models. It was consistently observed across all configurations that the 13B LLaMA-2 model outperformed the 7B LLaMA-2, corroborating the superior reasoning capabilities and more extensive implicit knowledge repository as discussed in [Section 4.3](#).

Significant impacts on performance were observed when captions were removed from the visual context, irrespective of the object detection model employed. For instance, the 13B configuration employing only DETIC achieved a 42.81% VQA score compared to 63.57% for the 13B configuration with both Caption and DETIC, a decline of 20.76%. Similar trends were observed in the EM scores with a comparable decline of 21.55%. This pronounced impact underscores the critical role of the well-designed prompt used with InstructBLIP, as outlined in [Appendix B.2](#), and it becomes particularly apparent by the less pronounced performance drop of 6.08% observed when object lists were removed instead of captions. This indicates that the visual context provided by captions is more influential than that provided by the object detection module. The lesser impact observed by the object detection module may also be ascribed to the limited requirement of the OK-VQA dataset for spatial comprehension of the visual scene, where the mapping of objects to their bounding boxes plays a crucial role in enhancing spatial awareness and overall model comprehension of the depicted scene.

Further observations from the experiments indicate that employing DETIC for object detection yielded better results, whether used with or without captions, in comparison to YOLOv5. For example, the 13B (Only DETIC) configuration performed 1.39% and 1.9% better than the 13B (Only YOLOv5) in VQA and EM scores, respectively. Although these improvements are marginal, they reinforce the assertion that a dataset demanding greater spatial awareness of the visual scene will exhibit a more pronounced difference between the use of DETIC and YOLOv5, due to DETIC's capability to detect a richer array of object classes compared to YOLOv5.

These findings highlight the nuanced contributions of various system components to the overall model efficacy and advocate for continued refinement and evaluation of these elements to bolster the performance of KB-VQA systems in complex visual environments.

<sup>17</sup> Due to the cost involved in using GPT-4, syntactic evaluation was chosen as the sole criterion for baselining in the ablation study.

## Chapter 5

# 5. Ethical Considerations

### 5.1. Limitations & Broader Impact

The Knowledge-Based Visual Question Answering (KB-VQA) system exhibits several critical limitations, including information loss during the image-to-language transformation, instances of incorrect reasoning that leads to misidentification of relevant objects, and challenges recognizing hazy or unclear text and telling time from clock images. These technical issues may potentially be exacerbated by inherent biases from the Pre-Trained Large Language Models (PT-LLMs) used, which can carry biases from their training data, affecting the fairness and neutrality of responses. Despite the effort made during this research to fine-tune the PT-LLM model with the OK-VQA dataset in an attempt to control its output, studies indicate that VQA datasets generally may still contain gender and racial biases that could lead to stereotyping (Hirota, Nakashima and Garcia, 2022). Additionally, the complexity of PT-LLMs often renders their decision-making processes less interpretable, raising significant concerns for trust and reliability in KB-VQA applications. Moreover, the necessity for continuous updates to keep these models relevant poses a challenge in rapidly evolving knowledge landscapes.

The deployment of KB-VQA systems enhances educational experiences by providing interactive and immediate answers to visual queries, democratizing access to information, and transforming learning environments to promote deeper understanding and retention of information. In the technological realm, the development of KB-VQA systems advances the fields of artificial intelligence, specifically in computer vision and natural language processing, pushing the boundaries of AI capabilities and leading to more sophisticated, efficient, and accurate systems applicable across various industries. Societally, KB-VQA systems improve the accuracy and responsiveness of AI tools, benefiting society by aiding the visually impaired, enhancing digital interactions, and providing crucial support in areas like medical imaging and emergency response. However, these systems also confront significant ethical concerns such as data privacy, security, and potential environmental impacts due to their computational demands.

Regulatory and policy considerations are becoming increasingly crucial as these technologies integrate more deeply into critical sectors. Despite achieving superior or comparable results, the inherent limitations, and broader impacts of the KB-VQA system necessitate cautious deployment. It is imperative to thoroughly assess the system to ensure its responsible implementation and alignment with broader societal needs and ethical standards.

### 5.2. Reproducibility

This research recognizes the critical importance of reproducibility in scientific inquiry and has implemented several measures to ensure the reproducibility of both the implementation and the results. To facilitate this, fixed seeds were used in all computations where possible to minimize variations in results due to random operations. Comprehensive details of the implementation, including all hyperparameters and settings, are thoroughly documented in [Appendix B](#). Additionally, the complete implementation code has been made

available on the HuggingFace Space detailed in [Section 3.3.4](#). These steps are taken to provide transparency and allow for the replication of the study's findings by other researchers, thereby contributing to the robustness and reliability of the research outcomes.

## Chapter 6

# 6. Conclusion & Future Work

### 6.1. Conclusion

This research delves into the Knowledge-Based Visual Question Answering (KB-VQA) domain, highlighting how visual questions often require external knowledge beyond what is depicted in the images to derive accurate answers. It provides a historical overview of the evolution of machine learning, placing particular emphasis on the revolutionary impact of the Transformer model, which has fundamentally advanced the field of language processing and multimodal tasks through its effective handling of sequential data. Additionally, the study explores mainstream Pre-Trained Large Language Models (PT-LLMs) and Pre-Trained Multimodal Models (PT-LMMs). These technologies have further shaped advancements in machine learning by leveraging vast, pre-trained knowledge bases to address complex tasks, demonstrating their pivotal role in the development and enhancement of KB-VQA systems. An extensive review of established methods for addressing the KB-VQA challenge has led to the adoption of a refined approach introduced in this research, which translates visual content into language space. This involves transforming images into detailed captions and lists of objects with their bounding boxes, leveraging the vast implicit knowledge and reasoning capabilities housed within PT-LLMs. The research refined the process of fine-tuning the PT-LLM by incorporating special tokens into the model's vocabulary to enhance its ability to interpret visual contexts effectively. It also analyzed current methods for image representation and sources of knowledge utilized in existing approaches, advocating for the use of implicit knowledge stored in PT-LLMs, particularly for tasks that do not require specialized knowledge.

Extensive ablation experiments assessed the impact of each component of the visual context on the overall model performance, emphasizing that the image descriptions generated during the captioning stage are the most crucial elements. Additionally, the study conducted a thorough analysis of mainstream KB-VQA datasets, with a particular focus on the OK-VQA dataset. Evaluation metrics were critically examined, detailing their strengths and weaknesses, and the evaluation process was enhanced by introducing semantic evaluation using GPT-4 to provide a more accurate understanding of performance aligned with real-world application needs.

Evaluation results demonstrate that the developed model achieves competent and competitive performance, recording a VQA score of 63.57% under the universal benchmark of syntactic evaluation, and excelling with an EM score of 68.36%. Further, semantic evaluations yielded even more impressive outcomes, with VQA and EM scores of 71.09% and 72.55%, respectively. These results indicate that the model effectively applies reasoning over the visual context and successfully retrieves the necessary knowledge to answer the visual questions.

Additionally, the findings underscore that the primary reasons for failure scenarios include information loss during the image-to-text transformation process, as well as the model's difficulties in recognizing text, brand logos, and time telling from clock images. Moreover, failures are further exacerbated by the inherent nature of the dataset, which includes questions with ambiguity and subjectivity, along with bias and inconsistency in the

ground truth answers. These elements collectively contribute to the challenges faced by the model in accurately interpreting and responding to the visual queries.

While the KB-VQA system developed in this research demonstrates superior performance on the EM score, it is important to recognize that not all models are assessed using this metric. Although the VQA score is broadly accepted as a standard benchmark, the EM score provides supplementary insights but cannot be regarded as a comprehensive standard for all models. The less stringent requirements of EM scores imply that models typically evaluated using only the VQA score might exceed existing benchmarks if also assessed using the EM score. Additionally, the existing evaluation methods, including both the VQA and EM scores, do not adequately capture the semantic accuracy of the responses provided by KB-VQA systems. Therefore, there is a clear need for a unified metric that incorporates the semantic meaning of answers. Implementing such a metric would enable a more comprehensive and accurate evaluation, ensuring a holistic and standardized approach to benchmarking that truly reflects the nuanced and precise capabilities of the KB-VQA models.

Overall, this research not only clarifies the complexities associated with KB-VQA but also sets the stage for future advancements in integrating visual content with advanced language models, thereby enhancing the interpretative and reasoning capabilities of AI systems in handling real-world visual queries.

## 6.2. Future Work

Although, the current approach of language-mediated visual context discussed in Chapter 3 has demonstrated exemplary performance, further explorations and investigations can be considered:

### 1. Multi-Modal Embeddings Alignment

The inception of this research was guided by two principal objectives: firstly, the immediate goal of devising a system to address the Knowledge-Based Visual Question Answering (KB-VQA) task via language-mediated visual context, which constitutes the primary focus of this dissertation and is detailed in Chapter 3; secondly, the long-term ambition to cultivate a versatile multimodal model capable of comprehensive vision and language understanding, as well as instruction following.

The preliminary architectural blueprint for this methodology draws inspiration from a spectrum of preceding studies (Alayrac et al., 2022; Merullo et al., 2023; Girdhar et al., 2023; R. Zhang et al., 2023; Dai et al., 2023; Li et al., 2023; Liu et al., 2023) and is depicted in Figure 6-1. The concept entails the generation of image embeddings via the pre-trained vision encoder of CLIP<sup>18</sup> (Radford et al., 2021), which are then fed into either a linear (Merullo et al., 2023; Liu et al., 2023) or attention-based (Alayrac et al., 2022; R. Zhang et al., 2023; Dai et al., 2023) projection layer. This layer is trained to predict corresponding caption textual embeddings produced by LLaMA-2, thereby aligning the image and textual embeddings into a unified space. Initially, all models are kept frozen

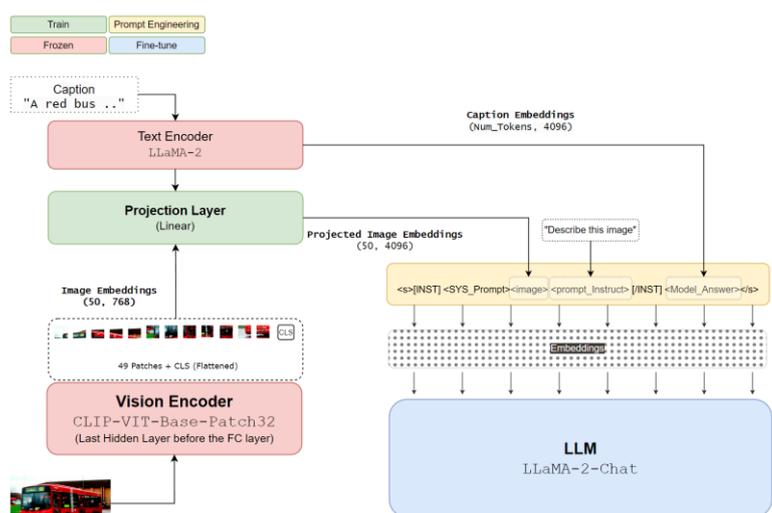


Figure 6-1 Blueprint design for Language-Vision embeddings alignment for Multimodal learning and Instruction-Following.

<sup>18</sup> CLIP has 12 transformer layers in the vision encoder followed by a fully connected layer. The embeddings are taken from the last transformer layer.

except the projection layer, which is trained on a large image-caption dataset. Subsequently, in the next training stage, both the projection layer and LLaMA-2, specifically LLaMA-2-Chat, are further trained on a selective high-quality image instruction-following dataset, while the vision encoder is kept frozen. The objective of the initial training stage is to convert image embeddings into visual tokens comprehensible by LLaMA-2, while the objective of the second training stage is to train LLaMA-2 on multi-modal instruction following.

Throughout the course of this dissertation, considerable time and effort were dedicated to this ambitious objective, exploring all available options for each component of the desired model. The potential for this approach to enhance a range of vision-related tasks beyond KB-VQA is compelling and warrants further investigation. Nonetheless, the substantial demands on computational resources and the requirement for massive image-text datasets for effective training render this approach currently impractical within the resource-limited context of this research. Future enhancements to the existing model under more favorable conditions will consider this innovative approach.

## 2. Dataset Enhancement

While the OK-VQA dataset stands as one of the premier resources for testing the reasoning capabilities of KB-VQA models, a comprehensive revision of this dataset is crucial, particularly with a focus on refining the ground truth answers. Preliminary efforts by (Reichman et al., 2023) have initiated this process; however, these modifications necessitate thorough validation to confirm their effectiveness. This revision is aimed at identifying and rectifying any inconsistencies, biases, or errors within the dataset, thereby enhancing the reliability and accuracy of the ground truth responses. It is essential that these improvements undergo methodical assessment through rigorous testing and cross-validation procedures to substantiate their impact on elevating the dataset's overall quality and utility in training more robust models. Moreover, the dataset could benefit from an expansion to include more samples that require spatial comprehension of the visual scene, thereby testing the model's ability to understand correlations between objects within the scene for better utilization of the provided object bounding boxes. Additionally, a critical component of this revision should include assessing the magnitude of bias within the dataset. This will involve analyzing the dataset for any inherent prejudicial elements that could skew model training and performance, ensuring that the models trained on this dataset operate fairly and effectively across diverse scenarios.

## 3. Enhancement of Component Models

The findings outlined in 4.3 highlight that larger Pre-Trained Large Language Models (PT-LLMs) demonstrate enhanced reasoning capabilities and possess a more comprehensive implicit knowledge base. Building on this, future studies could consider employing the 70B variant of LLaMA-2 to potentially boost performance, with a comparative analysis against the smaller variants previously utilized in experiments. This initiative aims to quantify improvements and further optimize model effectiveness.

Furthermore, as detailed in Section 4.5, the model faced significant challenges in text recognition within visual scenes. Addressing this, a promising direction for future research involves the integration of a sophisticated Optical Character Recognition (OCR) module. This enhancement would improve the model's processing of visual context, substantially enhancing its ability to accurately interpret and respond to images containing text.

Another innovative approach for future exploration is the incorporation of a "chain of thoughts" methodology (Wei et al., 2023) into the fine-tuning prompts. This technique would encourage the model to emulate a sequential, step-by-step reasoning process, potentially improving its handling of complex queries. This could be particularly beneficial for queries that demand spatial comprehension of the scene, enabling the model to produce more accurate and contextually relevant responses.

#### **4. Specialized Knowledge Application**

The KB-VQA model's tendency to underperform when faced with specialized knowledge questions presents an opportunity for targeted improvements through training in specific domains. Potential enhancements include deploying the model as an interactive guide in museums, where it could provide visitors with detailed explanations about exhibits in response to image-based queries. In educational settings, the model could be integrated into learning platforms to deliver in-depth information and explanations directly linked to visual content from textbooks or study materials. For the fashion and retail industry, the model could analyze images to offer fashion advice, check inventory levels, or facilitate virtual try-on experiences. Additionally, in the culinary field, the model could analyze dish images to provide recipes, nutritional facts, or cooking tips, enriching the user's culinary experience.

Furthermore, the model could also be adapted as assistive technology for the visually impaired, offering descriptive audio responses based on visual data to enhance accessibility and independence. In medical diagnostics, the model could assist healthcare professionals by analyzing medical images and providing preliminary diagnostic responses. However, applications in sensitive areas such as healthcare and assistive technologies will require meticulous testing and comprehensive evaluation to ensure accuracy, reliability, and adherence to ethical standards. These enhancements aim to leverage the model's capabilities to improve its accuracy and utility in responding to real-world, image-based queries across various domains.

# Bibliography

- Alammar, J., 2018. *The Illustrated Transformer* [Online]. Available from: <https://jalammar.github.io/illustrated-transformer/> [Accessed 21 November 2023].
- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A. and Simonyan, K., 2022. *Flamingo: a Visual Language Model for Few-Shot Learning* [Online]. arXiv. Available from: <http://arxiv.org/abs/2204.14198> [Accessed 26 April 2024].
- AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M. and Ghazvininejad, M., 2022. *A Review on Language Models as Knowledge Bases* [Online]. arXiv. Available from: <http://arxiv.org/abs/2204.06031> [Accessed 17 April 2024].
- Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J.H., Shafey, L.E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Yuanzhong, Zhang, Y., Abrego, G.H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C.A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A.C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilov, D., So, D.R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Yuhuai, Xu, K., Xu, Yunhan, Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S. and Wu, Yonghui, 2023. *PaLM 2 Technical Report* [Online]. arXiv. Available from: <http://arxiv.org/abs/2305.10403> [Accessed 5 November 2023].
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L. and Parikh, D., 2015. VQA: Visual Question Answering. *2015 IEEE International Conference on Computer Vision (ICCV)* [Online], 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile. Santiago, Chile: IEEE, pp.2425–2433. Available from: <https://doi.org/10.1109/ICCV.2015.279> [Accessed 20 August 2023].
- Bahdanau, D., Cho, K. and Bengio, Y., 2014. *Neural Machine Translation by Jointly Learning to Align and Translate* [Online]. arXiv. Available from: <http://arxiv.org/abs/1409.0473> [Accessed 16 November 2023].
- Baltrusaitis, T., Ahuja, C. and Morency, L.-P., 2018. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* [Online], 41(2), pp.423–443. Available from: <https://doi.org/10.1109/TPAMI.2018.2798607>.
- Bansal, N., 2023. *Transformer Models 101: Getting Started — Part 1. Medium* [Online]. Available from: <https://towardsdatascience.com/transformer-models-101-getting-started-part-1-b3a77ccfa14d> [Accessed 20 November 2023].
- Breiman, L., 2001. Random Forests. *Kluwer Academic Publishers* [Online]. Available from: [<https://link.springer.com/article/10.1023/A:1010933404324>](<https://link.springer.com/article/10.1023/A:1010933404324>) [Accessed 12 December 2023].
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D., 2020. *Language Models are Few-Shot Learners* [Online]. arXiv. Available from: <http://arxiv.org/abs/2005.14165> [Accessed 28 August 2023].
- Cao, B., Lin, H., Han, X. and Sun, L., 2023. The Life Cycle of Knowledge in Big Language Models: A Survey. *Machine Intelligence Research* [Online], 21(2), pp.217–238. Available from: <https://doi.org/10.1007/s11633-023-1416-x>.
- Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A. and Chang, M.-W., 2023. *Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions?* [Online]. arXiv. Available from: <http://arxiv.org/abs/2302.11713> [Accessed 17 March 2024].
- Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Zhang, H. and Gan, C., 2023. *See, Think, Confirm: Interactive Prompting Between Vision and Language Models for Knowledge-based Visual Reasoning* [Online]. arXiv. Available from: <http://arxiv.org/abs/2301.05226> [Accessed 29 August 2023].
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S. and Fiedel, N., 2022. *PaLM: Scaling Language Modeling with Pathways* [Online]. arXiv. Available from: <http://arxiv.org/abs/2204.02311> [Accessed 5 November 2023].
- Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S. and Amodei, D., 2017. *Deep reinforcement learning from human preferences* [Online]. arXiv. Available from: <http://arxiv.org/abs/1706.03741> [Accessed 15 April 2024].
- Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang,

- Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V. and Wei, J., 2022. *Scaling Instruction-Finetuned Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2210.11416> [Accessed 14 March 2024].
- Clark, K., Luong, M.-T., Le, Q.V. and Manning, C.D., 2020. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators* [Online]. arXiv. Available from: <http://arxiv.org/abs/2003.10555> [Accessed 2 November 2023].
- Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning* [Online], 20(3), pp.273–297. Available from: <https://doi.org/10.1007/BF00994018>.
- Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P. and Hoi, S., 2023. *InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning* [Online]. arXiv. Available from: <http://arxiv.org/abs/2305.06500> [Accessed 11 March 2024].
- Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei, 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* [Online], 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), Miami, FL. Miami, FL: IEEE, pp.248–255. Available from: <https://doi.org/10.1109/CVPR.2009.5206848> [Accessed 13 April 2024].
- Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L., 2023. *QLORA: Efficient Finetuning of Quantized LLMs*.
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* [Online]. arXiv. Available from: <http://arxiv.org/abs/1810.04805> [Accessed 28 August 2023].
- Ding, Y., Yu, J., Liu, B., Hu, Y., Cui, M. and Wu, Q., 2022. MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* [Online], 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA. New Orleans, LA, USA: IEEE, pp.5079–5088. Available from: <https://doi.org/10.1109/CVPR52688.2022.00503> [Accessed 29 August 2023].
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houshby, N., 2021. *AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE*.
- Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C., 2016. *The Importance of Skip Connections in Biomedical Image Segmentation* [Online]. arXiv. Available from: <http://arxiv.org/abs/1608.04117> [Accessed 21 November 2023].
- Du, Y., Li, J., Tang, T., Zhao, W.X. and Wen, J.-R., 2023. *Zero-shot Visual Question Answering with Language Model Feedback* [Online]. arXiv. Available from: <http://arxiv.org/abs/2305.17006> [Accessed 5 March 2024].
- Du, Y., Liu, Z., Li, J. and Zhao, W.X., 2022. *A Survey of Vision-Language Pre-Trained Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2202.10936> [Accessed 28 August 2023].
- de Faria, A.C.A.M., Bastos, F. de C., da Silva, J.V.N.A., Fabris, V.L., Uchoa, V. de S., Neto, D.G. de A. and Santos, C.F.G. dos, 2023. *Visual Question Answering: A Survey on Techniques and Common Trends in Recent Literature* [Online]. arXiv. Available from: <http://arxiv.org/abs/2305.11033> [Accessed 24 August 2023].
- Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D., 2008. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* [Online], 32(9), pp.1627–1645. Available from: <https://doi.org/10.1109/TPAMI.2009.167>.
- Gao, F., Ping, Q., Thattai, G., Reganti, A., Wu, Y.N. and Natarajan, P., 2022a. *A Thousand Words Are Worth More Than a Picture: Natural Language-Centric Outside-Knowledge Visual Question Answering* [Online]. arXiv. Available from: <http://arxiv.org/abs/2201.05299> [Accessed 1 September 2023].
- Gao, F., Ping, Q., Thattai, G., Reganti, A., Wu, Y.N. and Natarajan, P., 2022b. Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* [Online], 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA. New Orleans, LA, USA: IEEE, pp.5057–5067. Available from: <https://doi.org/10.1109/CVPR52688.2022.00501> [Accessed 18 March 2024].
- Gardères, F., Ziaeefard, M., Abeloos, B. and Lecue, F., 2020. ConceptBert: Concept-Aware Representation for Visual Question Answering. *Findings of the Association for Computational Linguistics: EMNLP 2020* [Online], Findings of the Association for Computational Linguistics: EMNLP 2020, Online. Online: Association for Computational Linguistics, pp.489–498. Available from: <https://doi.org/10.18653/v1/2020.findings-emnlp.44> [Accessed 9 September 2023].
- Geman, D., Geman, S., Hallonquist, N. and Younes, L., 2015. Visual Turing test for computer vision systems. *Proceedings of the National Academy of Sciences* [Online], 112(12), pp.3618–3623. Available from: <https://doi.org/10.1073/pnas.1422953112>.
- Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A. and Misra, I., 2023. *ImageBind: One Embedding Space To Bind Them All* [Online]. arXiv. Available from: <http://arxiv.org/abs/2305.05665> [Accessed 26 April 2024].
- Google, 2023. *Introducing Gemini*. Google [Online]. Available from: <https://blog.google/technology/ai/google-gemini-ai/> [Accessed 11 March 2024].
- Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. and Parikh, D., 2017. *Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering* [Online]. arXiv. Available from: <http://arxiv.org/abs/1612.00837> [Accessed 28 April 2024].
- Graves, A., 2013. *Generating Sequences With Recurrent Neural Networks* [Online]. arXiv. Available from: <http://arxiv.org/abs/1308.0850> [Accessed 17 November 2023].
- Gui, L., Wang, B., Huang, Q., Hauptmann, A., Bisk, Y. and Gao, J., 2022. *KAT: A Knowledge Augmented Transformer for Vision-and-Language* [Online]. arXiv. Available from: <http://arxiv.org/abs/2112.08614> [Accessed 29 August 2023].
- Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D. and Hoi, S.C.H., 2023. *From Images to Textual Prompts: Zero-shot VQA with Frozen Large Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2212.10846> [Accessed 30 April 2024].
- Han, X., Wang, Y.-T., Feng, J.-L., Deng, C., Chen, Z.-H., Huang, Y.-A., Su, H., Hu, L. and Hu, P.-W., 2023. A survey of transformer-based multimodal pre-trained modals. *Neurocomputing* [Online], 515, pp.89–106. Available from: <https://doi.org/10.1016/j.neucom.2022.09.136>.

- Han, Xiaotian, Yang, J., Hu, H., Zhang, L., Gao, J. and Zhang, P., 2021. *Image Scene Graph Generation (SGG) Benchmark* [Online]. arXiv. Available from: <http://arxiv.org/abs/2107.12604> [Accessed 10 March 2024].
- Han, Xu, Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.-R., Yuan, J., Zhao, W.X. and Zhu, J., 2021. Pre-trained models: Past, present and future. *AI Open* [Online], 2, pp.225–250. Available from: <https://doi.org/10.1016/j.aiopen.2021.08.002>.
- He, P., Liu, X., Gao, J. and Chen, W., 2021. *DeBERTa: Decoding-enhanced BERT with Disentangled Attention* [Online]. arXiv. Available from: <http://arxiv.org/abs/2006.03654> [Accessed 2 November 2023].
- Hirota, Y., Nakashima, Y. and Garcia, N., 2022. Gender and Racial Bias in Visual Question Answering Datasets. *2022 ACM Conference on Fairness, Accountability, and Transparency* [Online], FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea. Seoul Republic of Korea: ACM, pp.1280–1292. Available from: <https://doi.org/10.1145/3531146.3533184> [Accessed 4 May 2024].
- Howard, J. and Ruder, S., 2018. *Universal Language Model Fine-tuning for Text Classification* [Online]. arXiv. Available from: <http://arxiv.org/abs/1801.06146> [Accessed 28 August 2023].
- Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., 2021. *LoRA: Low-Rank Adaptation of Large Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2106.09685> [Accessed 21 April 2024].
- Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D.L. and Szekely, P., 2021. *Dimensions of Commonsense Knowledge* [Online]. arXiv. Available from: <http://arxiv.org/abs/2101.04640> [Accessed 6 March 2024].
- Itti, L. and Koch, C., 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* [Online], 2(3), pp.194–203. Available from: <https://doi.org/10.1038/35058500>.
- Jocher, G., 2020. *YOLOv5* [Online]. Available from: <https://github.com/ultralytics/yolov5>.
- Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D. and Jawahar, C.V., 2021. *MMBERT: Multimodal BERT Pretraining for Improved Medical VQA* [Online]. arXiv. Available from: <http://arxiv.org/abs/2104.01394> [Accessed 14 March 2024].
- Kim, C., 2019. *Skip Connections and Residual Blocks* [Online]. Available from: <https://christina.kim/2019/10/29/residual-blocks-and-skip-connections/> [Accessed 21 November 2023].
- Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* [Online], 60(6), pp.84–90. Available from: <https://doi.org/10.1145/3065386>.
- LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature* [Online], 521(7553), pp.436–444. Available from: <https://doi.org/10.1038/nature14539>.
- Li, J., Li, D., Savarese, S. and Hoi, S., 2023. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2301.12597> [Accessed 11 March 2024].
- Li, J., Li, D., Xiong, C. and Hoi, S., 2022. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation* [Online]. arXiv. Available from: <http://arxiv.org/abs/2201.12086> [Accessed 10 March 2024].
- Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W. and Gao, J., 2022. *Grounded Language-Image Pre-training* [Online]. arXiv. Available from: <http://arxiv.org/abs/2112.03857> [Accessed 5 March 2024].
- Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y. and Gao, J., 2020. *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks* [Online]. arXiv. Available from: <http://arxiv.org/abs/2004.06165> [Accessed 10 March 2024].
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L. and Dollár, P., 2014. *Microsoft COCO: Common Objects in Context* [Online]. arXiv. Available from: <http://arxiv.org/abs/1405.0312> [Accessed 30 August 2023].
- Lin, Y., Xie, Y., Chen, D., Xu, Y., Zhu, C. and Yuan, L., 2022. *REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering*.
- Liu, H., Li, C., Wu, Q. and Lee, Y.J., 2023. *Visual Instruction Tuning* [Online]. arXiv. Available from: <http://arxiv.org/abs/2304.08485> [Accessed 11 March 2024].
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach* [Online]. arXiv. Available from: <http://arxiv.org/abs/1907.11692> [Accessed 2 November 2023].
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* [Online], 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada. Montreal, QC, Canada: IEEE, pp.9992–10002. Available from: <https://doi.org/10.1109/ICCV48922.2021.00986> [Accessed 28 August 2023].
- Lu, J., Batra, D., Parikh, D. and Lee, S., 2019. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P. and Kalyan, A., 2022. *Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering* [Online]. arXiv. Available from: <http://arxiv.org/abs/2209.09513> [Accessed 14 March 2024].
- Luong, M.-T., Pham, H. and Manning, C.D., 2015. *Effective Approaches to Attention-based Neural Machine Translation* [Online]. arXiv. Available from: <http://arxiv.org/abs/1508.04025> [Accessed 17 November 2023].
- Lymperaio, M. and Stamou, G., 2023. *A survey on knowledge-enhanced multimodal learning* [Online]. arXiv. Available from: <http://arxiv.org/abs/2211.12328> [Accessed 27 August 2023].
- Ma, Z., Li, S., Sun, B., Cai, J., Long, Z. and Ma, F., 2024. *GeReA\_Question\_Aware\_Prompt\_Captions\_for\_Knowledge\_based\_Visual\_Question\_Answering*.

- Mañas, O., Krojer, B. and Agrawal, A., 2024. *Improving Automatic VQA Evaluation Using Large Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2310.02567> [Accessed 26 March 2024].
- Manmadhan, S. and Kooor, B.C., 2020. Visual question answering: a state-of-the-art review. *Artificial Intelligence Review* [Online], 53(8), pp.5705–5745. Available from: <https://doi.org/10.1007/s10462-020-09832-7>.
- Manning, C. and Schütze, H., 1999. Foundations of statistical natural language processing. *ACM SIGMOD Record* [Online], 31(3), pp.37–38. Available from: <https://doi.org/10.1145/601858.601867>.
- Marino, K., Chen, X., Parikh, D., Gupta, A. and Rohrbach, M., 2021. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* [Online], 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. Nashville, TN, USA: IEEE, pp.14106–14116. Available from: <https://doi.org/10.1109/CVPR46437.2021.01389> [Accessed 29 August 2023].
- Marino, K., Rastegari, M., Farhadi, A. and Mottaghi, R., 2019. *OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge* [Online]. arXiv. Available from: <http://arxiv.org/abs/1906.00067> [Accessed 26 August 2023].
- Merullo, J., Castricato, L., Eickhoff, C. and Pavlick, E., 2023. *LINEARLY MAPPING FROM IMAGE TO TEXT SPACE*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. *Distributed Representations of Words and Phrases and their Compositionality* [Online]. Available from: [Accessed 2 November 2023].
- Minsky, M. and Papert, S.A., 1969. *Perceptrons: An Introduction to Computational Geometry* [Online]. The MIT Press. Available from: <https://doi.org/10.7551/mitpress/11301.001.0001> [Accessed 13 March 2024].
- Mnih, V., Heess, N. and Graves, A., 2014. *Recurrent Models of Visual Attention*.
- Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A., 2023. *A Comprehensive Overview of Large Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2307.06435> [Accessed 2 November 2023].
- Ng, A.Y. and Jordan, M.I., 2001. *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes*.
- OpenAI, 2023. *GPT-4 Technical Report* [Online]. arXiv. Available from: <http://arxiv.org/abs/2303.08774> [Accessed 29 August 2023].
- OpenAI, 2024a. *GPT-4 API Pricing* [Online]. Available from: <https://openai.com/pricing> [Accessed 26 March 2024].
- OpenAI, 2024b. *OpenAI Platform* [Online]. Available from: <https://platform.openai.com> [Accessed 27 March 2024].
- Orhan, A.E. and Pitkow, X., 2018. *Skip Connections Eliminate Singularities* [Online]. arXiv. Available from: <http://arxiv.org/abs/1701.09175> [Accessed 21 November 2023].
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E. and Launay, J., 2023. *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only* [Online]. arXiv. Available from: <http://arxiv.org/abs/2306.01116> [Accessed 5 November 2023].
- Pennington, J., Socher, R. and Manning, C., 2014. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [Online], Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. Doha, Qatar: Association for Computational Linguistics, pp.1532–1543. Available from: <https://doi.org/10.3115/v1/D14-1162> [Accessed 3 November 2023].
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018. *Deep contextualized word representations* [Online]. arXiv. Available from: <http://arxiv.org/abs/1802.05365> [Accessed 28 August 2023].
- Posner, M.I., 1980. Orienting of Attention. *Quarterly Journal of Experimental Psychology* [Online], 32(1), pp.3–25. Available from: <https://doi.org/10.1080/00335558008248231>.
- Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N. and Huang, X., 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* [Online], 63(10), pp.1872–1897. Available from: <https://doi.org/10.1007/s11431-020-1647-3>.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I., 2021. *Learning Transferable Visual Models From Natural Language Supervision*.
- Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. *Improving Language Understanding by Generative Pre-Training*.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. *Language Models are Unsupervised Multitask Learners*.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2019. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* [Online]. arXiv. Available from: <http://arxiv.org/abs/1910.10683> [Accessed 28 August 2023].
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I., 2021. *Zero-Shot Text-to-Image Generation* [Online]. arXiv. Available from: <http://arxiv.org/abs/2102.12092> [Accessed 29 August 2023].
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H., 2016. *Generative Adversarial Text to Image Synthesis* [Online]. arXiv. Available from: <http://arxiv.org/abs/1605.05396> [Accessed 6 May 2024].
- Reichman, B.Z., Sundar, A., Richardson, C., Zubatiy, T., Chowdhury, P., Shah, A., Truxal, J., Grimes, M., Shah, D., Chee, W.J., Punjwani, S., Jain, A. and Heck, L., 2023. Outside Knowledge Visual Question Answering Version 2.0. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [Online], ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece. Rhodes Island, Greece: IEEE, pp.1–5. Available from: <https://doi.org/10.1109/ICASSP49357.2023.10096074> [Accessed 17 August 2023].
- Ren, S., He, K., Girshick, R. and Sun, J., 2016. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks* [Online]. arXiv. Available from: <http://arxiv.org/abs/1506.01497> [Accessed 1 February 2024].
- Reynolds, L. and McDonnell, K., 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* [Online], CHI '21: CHI Conference on Human Factors in Computing Systems, Yokohama Japan. Yokohama Japan: ACM, pp.1–7. Available from: <https://doi.org/10.1145/3411763.3451760> [Accessed 16 April 2024].

- Rogge, N., 2023. *Comparing Captioning Models - a Hugging Face Space by nielsr* [Online]. Available from: <https://huggingface.co/spaces/nielsr/comparing-captioning-models> [Accessed 12 August 2023].
- Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* [Online], 65(6), pp.386–408. Available from: <https://doi.org/10.1037/h0042519>.
- Ruiz, A., 2023. *Why Llama 2 is a Game Changer. nocode.ai* [Online]. Available from: <https://www.nocode.ai/why-llama-2-is-game-changer/> [Accessed 15 April 2024].
- Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. *Learning representations by back-propagating errors* [Online]. Available from: <https://www.nature.com/articles/323533a0> [Accessed 12 December 2023].
- Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., Tow, J., Rush, A.M., Biderman, S., Webson, A., Ammanamanchi, P.S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A.V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P.O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A.F., Alfassy, A., Rogers, A., Nitzav, A.K., Xu, Canwen, Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D.I., Radev, D., Ponferrada, E.G., Levkovich, E., Kim, E., Natan, E.B., De Toni, F., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Froberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Von Werra, L., Weber, L., Phan, L., allal, L.B., Tanguy, L., Dey, M., Muñoz, M.R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M.T.-J., Vu, M.C., Jauhar, M.A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R.L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S.H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T.T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laipala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D.E., Salesky, E., Mielke, S.J., Lee, W.Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J.A., Rozen, J., Gao, L., Sutawika, L., Bari, M.S., Al-shaibani, M.S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S.H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojari, H., Roberts, A., Chung, H.W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P.F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Revena, S., Patil, S., Dettmers, T., Barua, A., Singh, Amanpreet, Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névól, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G.I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J.Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, Rui, Zhang, Ruochen, Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C.M., McDuff, D., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D.A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J.B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynek, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, Thanh, Oyebeade, T., Le, Trieu, Yang, Y., Nguyen, Z., Kashyap, A.R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, Ayush, Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, Chuxin, Fourier, C., Perifán, D.L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrmann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H.U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J.D., Sivaraman, K.R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M.H., Takeuchi, M., Pàmies, M., Castillo, M.A., Nezhurina, M., Sängler, M., Samwald, M., Cullan, M., Weinberg, M., De Wolf, M., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N.M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, Rosaline, Su, Ruisi, Cahyawijaya, S., Garda, S., Deshmukh, S.S., Mishra, S., Kiblawi, S., Ott, S., Sang-aaronsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y.S., Venkatraman, Y., Xu, Yifan, Xu, Yingxin, Xu, Yu, Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y. and Wolf, T., 2023. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model* [Online]. arXiv. Available from: <http://arxiv.org/abs/2211.05100> [Accessed 5 November 2023].
- Schwenk, D., Khandelwal, A., Clark, C., Marino, K. and Mottaghi, R., 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella and T. Hassner, eds. *Computer Vision – ECCV 2022* [Online], 13668. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp.146–162. Available from: [https://doi.org/10.1007/978-3-031-20074-8\\_9](https://doi.org/10.1007/978-3-031-20074-8_9) [Accessed 31 August 2023].
- Shao, Z., Yu, Z., Wang, M. and Yu, J., 2023. *Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering* [Online]. arXiv. Available from: <http://arxiv.org/abs/2303.01903> [Accessed 18 August 2023].
- Sharma, P., Ding, N., Goodman, S. and Soricut, R., 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* [Online], Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia. Melbourne, Australia: Association for Computational Linguistics, pp.2556–2565. Available from: <https://doi.org/10.18653/v1/P18-1238> [Accessed 13 April 2024].
- Si, Q., Mo, Y., Lin, Z., Ji, H. and Wang, W., 2023. *Combo of Thinking and Observing for Outside-Knowledge VQA*.
- Speer, R., Chin, J. and Havasi, C., 2017. *ConceptNet 5.5: An Open Multilingual Graph of General Knowledge* [Online]. arXiv. Available from: <http://arxiv.org/abs/1612.03975> [Accessed 29 August 2023].
- Stoelinga, S., 2023. *Calculating GPU memory for serving LLMs* [Online]. Available from: <https://www.substratus.ai/blog/calculating-gpu-memory-for-llm> [Accessed 21 April 2024].
- Sutskever, I., Vinyals, O. and Le, Q.V., 2014. *Sequence to Sequence Learning with Neural Networks* [Online]. Available from: [Accessed 5 January 2024].

- Sutton, R.S. and Barto, A.G., 1998. *Reinforcement Learning: An Introduction*.
- Tan, A.D.J. and Shen, B., 2023. *Tackling VQA with Pretrained Foundation Models without Further Training* [Online]. arXiv. Available from: <http://arxiv.org/abs/2309.15487> [Accessed 6 March 2024].
- Tan, H. and Bansal, M., 2019. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers* [Online]. arXiv. Available from: <http://arxiv.org/abs/1908.07490> [Accessed 14 March 2024].
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S. and Scialom, T., 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2307.09288> [Accessed 5 November 2023].
- Turing, A.M., 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* [Online], LIX(236), pp.433–460. Available from: <https://doi.org/10.1093/mind/LIX.236.433>.
- Ul Abideen, Z., 2023. *From Seq2Seq to Attention: Revolutionizing Sequence Modeling* [Online], June. Available from: <https://medium.com/@zaiinn440/from-seq2seq-to-attention-revolutionizing-sequence-modeling-67282ba82e83>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. *Attention is All you Need*.
- Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., 2015. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [Online], 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, Boston, MA, USA: IEEE, pp.3156–3164. Available from: <https://doi.org/10.1109/CVPR.2015.7298935> [Accessed 30 August 2023].
- Voita, E., Talbot, D., Moiseev, F., Sennrich, R. and Titov, I., 2019. *Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned* [Online]. arXiv. Available from: <http://arxiv.org/abs/1905.09418> [Accessed 20 November 2023].
- Wang, H. and Ge, W., 2024. *Q&A Prompts: Discovering Rich Visual Clues through Mining Question-Answer Prompts for VQA requiring Diverse World Knowledge* [Online]. arXiv. Available from: <http://arxiv.org/abs/2401.10712> [Accessed 10 March 2024].
- Wang, P., Wu, Q., Shen, C., Dick, A. and Van Den Hengel, A., 2015. Explicit Knowledge-based Reasoning for Visual Question Answering. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* [Online], Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, pp.1290–1296. Available from: <https://doi.org/10.24963/ijcai.2017/179> [Accessed 31 August 2023].
- Wang, P., Wu, Q., Shen, C., Hengel, A. van den and Dick, A., 2017. *FVQA: Fact-based Visual Question Answering* [Online]. arXiv. Available from: <http://arxiv.org/abs/1606.05433> [Accessed 29 August 2023].
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J. and Yang, H., 2022. *OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework* [Online]. arXiv. Available from: <http://arxiv.org/abs/2202.03052> [Accessed 28 April 2024].
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. and Zhou, D., 2023. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2201.11903> [Accessed 14 April 2024].
- Wikipedia, the free encyclopedia* [Online], n.d. Available from: <https://www.wikipedia.org/> [Accessed 7 January 2024].
- Wu, J., Lu, J., Sabharwal, A. and Mottaghi, R., 2021. *Multi-Modal Answer Validation for Knowledge-Based VQA*.
- Wu, Q., Wang, P., Wang, X., He, X. and Zhu, W., 2022. *Visual Question Answering: From Theory to Application* [Online]. Advances in Computer Vision and Pattern Recognition. Singapore: Springer Nature Singapore. Available from: <https://doi.org/10.1007/978-981-19-0964-1> [Accessed 27 August 2023].
- Xenos, A., Stafylakis, T., Patras, I. and Tzimiropoulos, G., 2023. *A Simple Baseline for Knowledge-Based Visual Question Answering* [Online]. arXiv. Available from: <http://arxiv.org/abs/2310.13570> [Accessed 14 April 2024].
- Xu, J., Sun, X., Zhang, Z., Zhao, G. and Lin, J., 2019. Understanding and Improving Layer Normalization. *Advances in Neural Information Processing Systems* [Online], 32. Curran Associates, Inc. Available from: [https://proceedings.neurips.cc/paper\\_files/paper/2019/hash/2f4fe03d77724a7217006e5d16728874-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2019/hash/2f4fe03d77724a7217006e5d16728874-Abstract.html) [Accessed 21 November 2023].
- Xu, P., Zhu, X. and Clifton, D.A., 2023. *Multimodal Learning with Transformers: A Survey* [Online]. arXiv. Available from: <http://arxiv.org/abs/2206.06488> [Accessed 20 August 2023].
- Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z. and Wang, L., 2022. *An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA* [Online]. arXiv. Available from: <http://arxiv.org/abs/2109.05014> [Accessed 29 August 2023].
- Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T. and Chen, E., 2023. *A Survey on Multimodal Large Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2306.13549> [Accessed 11 March 2024].
- Yu, Z., Yu, J., Cui, Y., Tao, D. and Tian, Q., 2019. *Deep Modular Co-Attention Networks for Visual Question Answering* [Online]. arXiv. Available from: <http://arxiv.org/abs/1906.10770> [Accessed 6 March 2024].
- Zakari, R.Y., Owusu, J.W., Wang, H., Qin, K., Lawal, Z.K. and Dong, Y., 2022. *VQA and Visual Reasoning: An Overview of Recent Datasets, Methods and Challenges* [Online]. arXiv. Available from: <http://arxiv.org/abs/2212.13296> [Accessed 17 March 2024].
- Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y. and Gao, J., 2021. *VinVL: Revisiting Visual Representations in Vision-Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2101.00529> [Accessed 5 March 2024].
- Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H. and Qiao, Y., 2023. *LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention* [Online]. arXiv. Available from: <http://arxiv.org/abs/2303.16199> [Accessed 26 April 2024].

- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T. and Zettlemoyer, L., 2022. *OPT: Open Pre-trained Transformer Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2205.01068> [Accessed 10 March 2024].
- Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., Guo, Y. and Zhang, L., 2023. *Recognize Anything: A Strong Image Tagging Model* [Online]. arXiv. Available from: <http://arxiv.org/abs/2306.03514> [Accessed 5 March 2024].
- Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y. and Wen, J.-R., 2023. *A Survey of Large Language Models* [Online]. arXiv. Available from: <http://arxiv.org/abs/2303.18223> [Accessed 2 November 2023].
- Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. and Gao, J., 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence* [Online], 34(07), pp.13041–13049. Available from: <https://doi.org/10.1609/aaai.v34i07.7005>.
- Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P. and Misra, I., 2022. *Detecting Twenty-thousand Classes using Image-level Supervision* [Online]. arXiv. Available from: <http://arxiv.org/abs/2201.02605> [Accessed 1 February 2024].
- Zhu, Z., Xue, Y., Chen, X., Zhou, D., Tang, J., Schuurmans, D. and Dai, H., 2023. *Large Language Models can Learn Rules* [Online]. arXiv. Available from: <http://arxiv.org/abs/2310.07064> [Accessed 14 April 2024].

## Chapter 7

# 7. Appendix

## A. Design

### A.1. YOLOv5 Detectable Object Classes

Below is the list of object classes detectable by the pretrained YOLOv5 (Jocher, 2020):

```
['person', 'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train',  
'truck', 'boat', 'traffic light', 'fire hydrant', 'stop sign', 'parking  
meter', 'bench', 'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant',  
'bear', 'zebra', 'giraffe', 'backpack', 'umbrella', 'handbag', 'tie',  
'suitcase', 'frisbee', 'skis', 'snowboard', 'sports ball', 'kite', 'baseball  
bat', 'baseball glove', 'skateboard', 'surfboard', 'tennis racket', 'bottle',  
'wine glass', 'cup', 'fork', 'knife', 'spoon', 'bowl', 'banana', 'apple',  
'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut',  
'cake', 'chair', 'couch', 'potted plant', 'bed', 'dining table', 'toilet',  
'tv', 'laptop', 'mouse', 'remote', 'keyboard', 'cell phone', 'microwave',  
'oven', 'toaster', 'sink', 'refrigerator', 'book', 'clock', 'vase',  
'scissors', 'teddy bear', 'hair drier', 'toothbrush']
```

## A.2. Token Distribution for Fine-tuning Data

During data preparation for LLaMA-2 fine-tuning the samples with token count exceeding 1,024 tokens were removed in order to minimize the computational requirements, this does not impact the fine-tuning process, because the fine-tuning objective is to learn how to leverage the visual context and deduce the correct prediction in adherence to the OK-VQA dataset structure. Below histogram shows the fine-tuning samples distribution with respect to token counts, before and after the filtration.

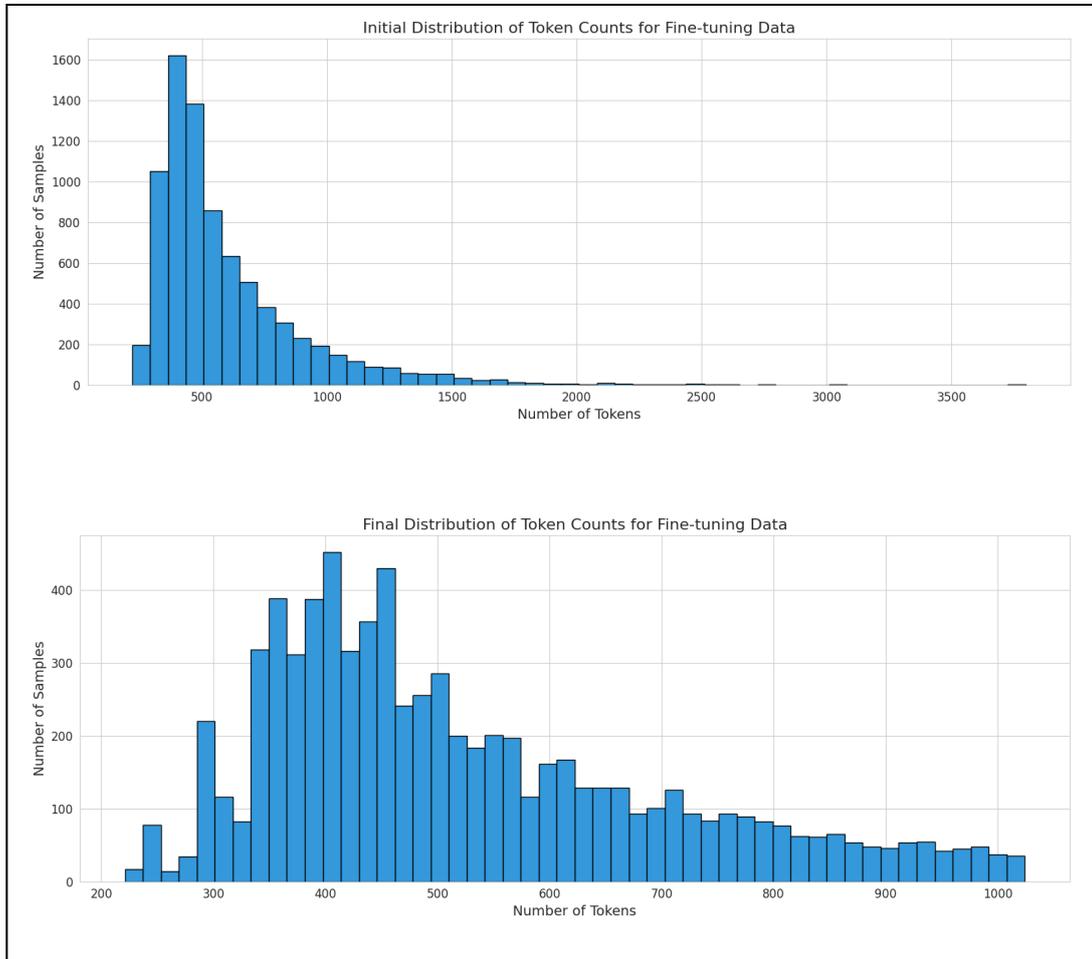


Figure 7-1 Token count distribution for the finetuning data before and after removing samples with more than 1024 tokens.

## A.3. Default LLaMA-2 System Prompt

Below is the default LLaMA-2 Chat system prompt (Touvron et al., 2023):

```
You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.
```

### A.4. KB-VQA System Prompt

The following system prompt was designed and used for fine-tuning, evaluation, and inference of the KB-VQA model:

```
"You are a helpful, respectful, and honest assistant for visual question answering. you are provided with a caption of an image and a list of objects detected in the image along with their bounding boxes and level of certainty, you will output an answer to the given questions in no more than one sentence. Use logical reasoning to reach to the answer, but do not output your reasoning process unless asked for it. If provided, you will use the [CAP] and [/CAP] tags to indicate the beginning and end of the caption respectively. If provided you will use the [OBJ] and [/OBJ] tags to indicate the beginning and end of the list of detected objects in the image along with their bounding boxes respectively. If provided, you will use [QES] and [/QES] tags to indicate the beginning and end of the question respectively."
```

### A.5. Comparison of Captioning Models

Below is a screenshot of the HuggingFace space used during assessment of various multimodal models for image captioning.

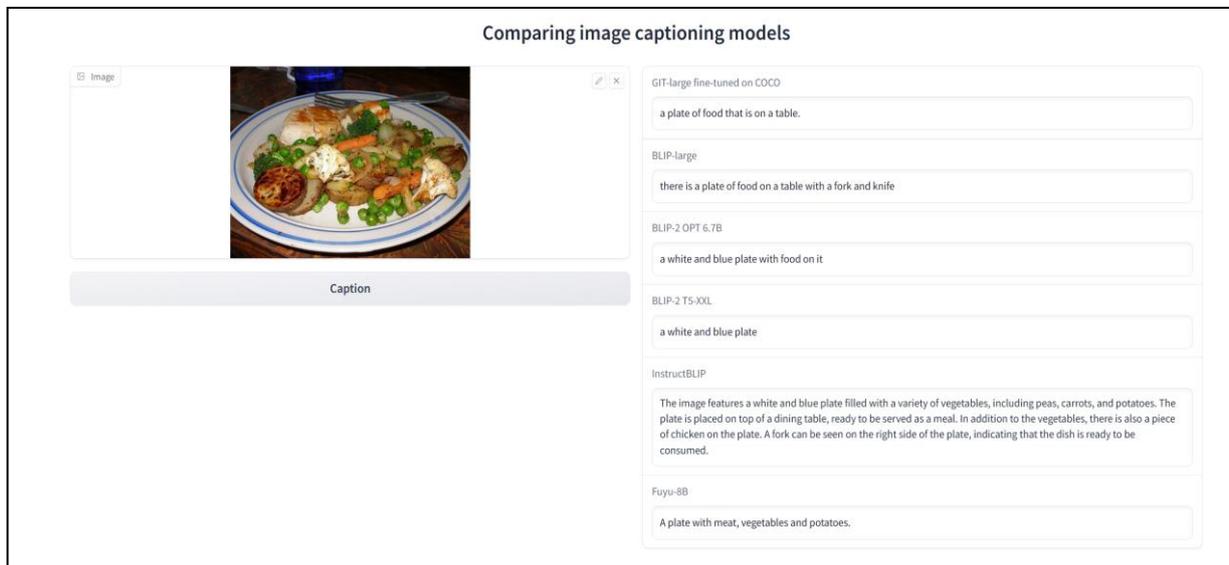


Figure 7-2 Comparison between various captioning models demonstrating InstructBLIP superiority.

## B. Implementation

### B.1. Overall Implementation Flow

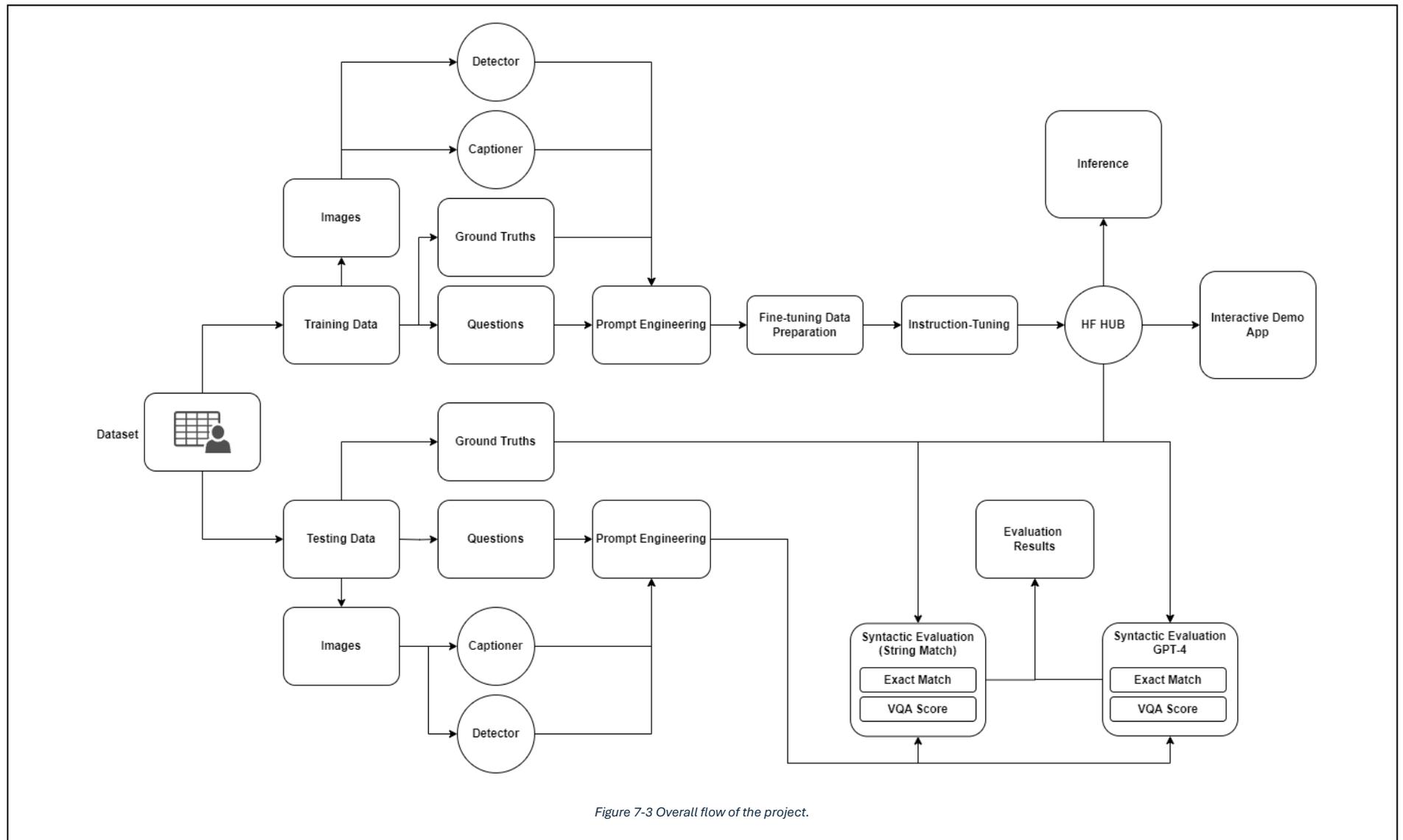


Figure 7-3 Overall flow of the project.

## B.2. Component Models Configurations and Hyperparameters

Component Models		
Captioner	MODEL_NAME	Salesforce/instructblip-vicuna-7b
	MAX_IMAGE_SIZE	1024
	MIN_LENGTH	150
	MAX_NEW_TOKENS	400
	QUANTIZATION	4bit
	TORCH_DTYPE	torch.float16
	LOW_CPU_MEM_USAGE	TRUE
PROMPT	<pre>"Provide a comprehensive and detailed description of the following image. Focus on identifying and describing every element in the scene, including all people (along with their gender, age, colour, and any prominent feature), animals along with their breed and all objects, their count, their positions, and any actions or interactions taking place. Pay special attention to the positioning of limbs and hands, and any objects they might be holding or interacting with. Describe texts, colours, textures, setting, atmosphere, mood, and any indicators of the time of day, such as the quality of light, shadows. Ensure to capture both the obvious and subtle elements for a complete understanding of the image. Answer as if you were looking at the image."</pre>	
PT-LLM	MODEL_NAME	meta-llama/Llama-2-7b-chat-hf meta-llama/Llama-2-13b-chat-hf
	QUANTIZATION	4bit
	USE_FAST_TOKENIZER	TRUE
Detector	MODEL_NAME	ultralytics/yolov5 deformable-detr-detic
	CONFIDENCE_THRESHOLD	0.2

Table 7-1 Component models configurations and hyperparameters

## B.3. Fine-tuning configurations and Hyperparameters

Fine-tuning Configurations	
NUM_TRAIN_EPOCHS	1
GRADIENT_ACCUMULATION_STEPS	1
OPTIMIZER	paged_adamw_8bit
LEARNING_RATE	0.0002
WEIGHT_DECAY	0.01
EVALUATION_STRATEGY	steps
EVALUATION_STEPS	5
MAX_GRAD_NORM	0.3
LR_SCHEDULER_TYPE	linear
TARGET_MODULES	['up_proj', 'down_proj', 'k_proj', 'q_proj', 'v_proj', 'o_proj']
FP16	TRUE
PER_DEVICE_TRAIN_BATCH_SIZE	16
PER_DEVICE_EVAL_BATCH_SIZE	8
GROUP_BY_LENGTH	FALSE
MAX_TOKEN_COUNT	1024
PACKING	FALSE
TEST_SIZE	0.1
SEED	123
WARMUP_RATIO	0.03
FP16	TRUE
LORA_R	64
LORA_ALPHA	32
LORA_DROPOUT	0.05

Table 7-2 Fine-tuning configurations and hyperparameters.

### B.4. Fine-tuning Results

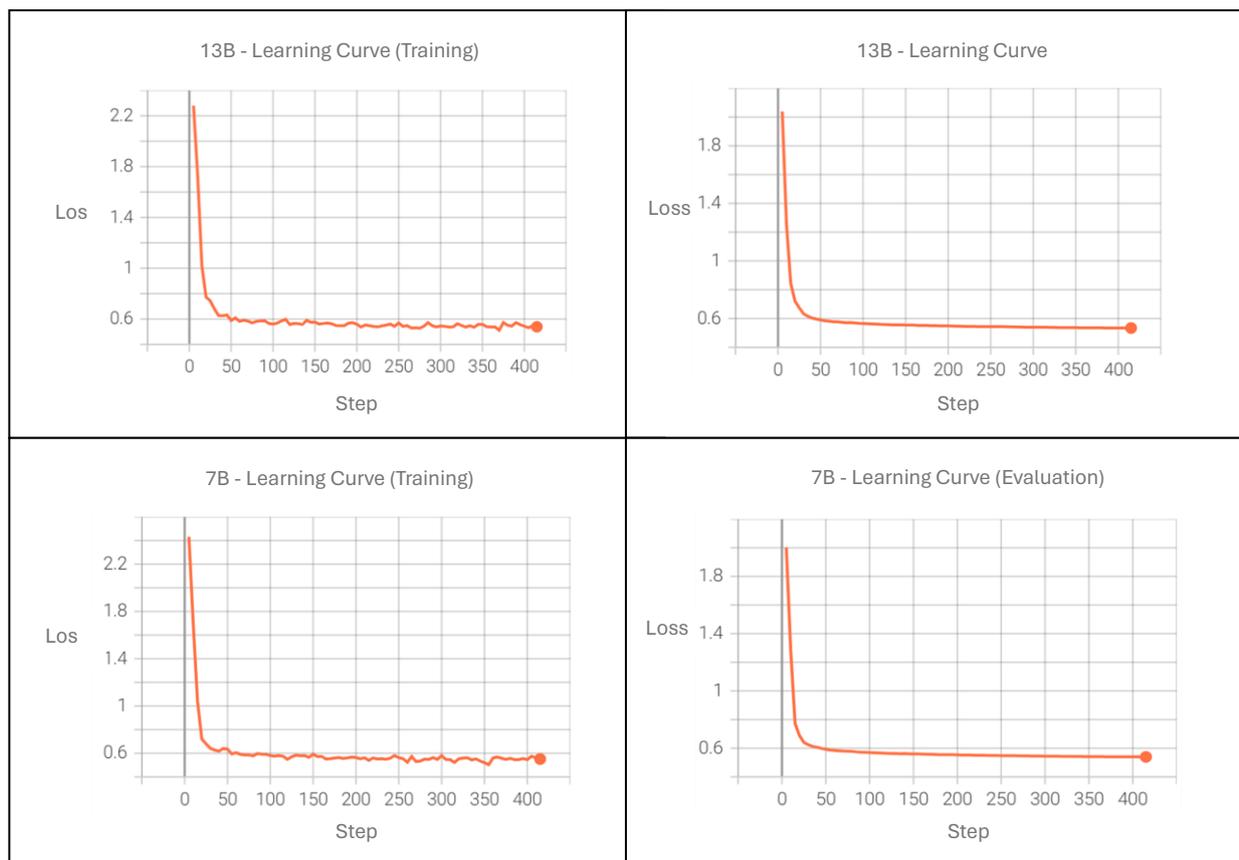


Figure 7-4 Fine-tuning learning curves.

## C. Evaluation

### C.1. Token Count for Evaluation Data (DETIC vs YOLOv5)

Below scatterplot show the number of tokens distribution for the evaluation data when employing DETIC and YOLOv5 for objects detection

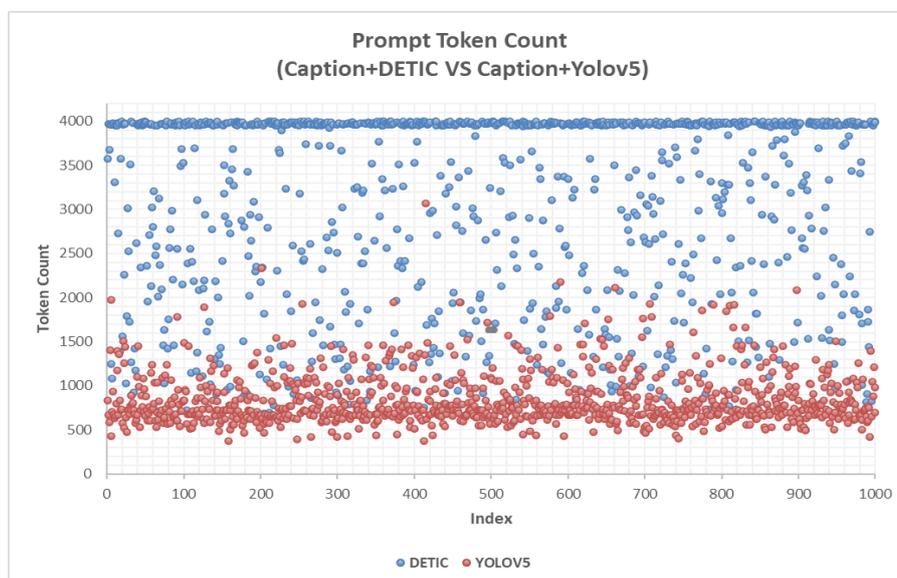


Figure 7-5 Token Count for Evaluation Data (DETIC vs YOLOv5)

## C.2. Ablation Study Scores (Graph)

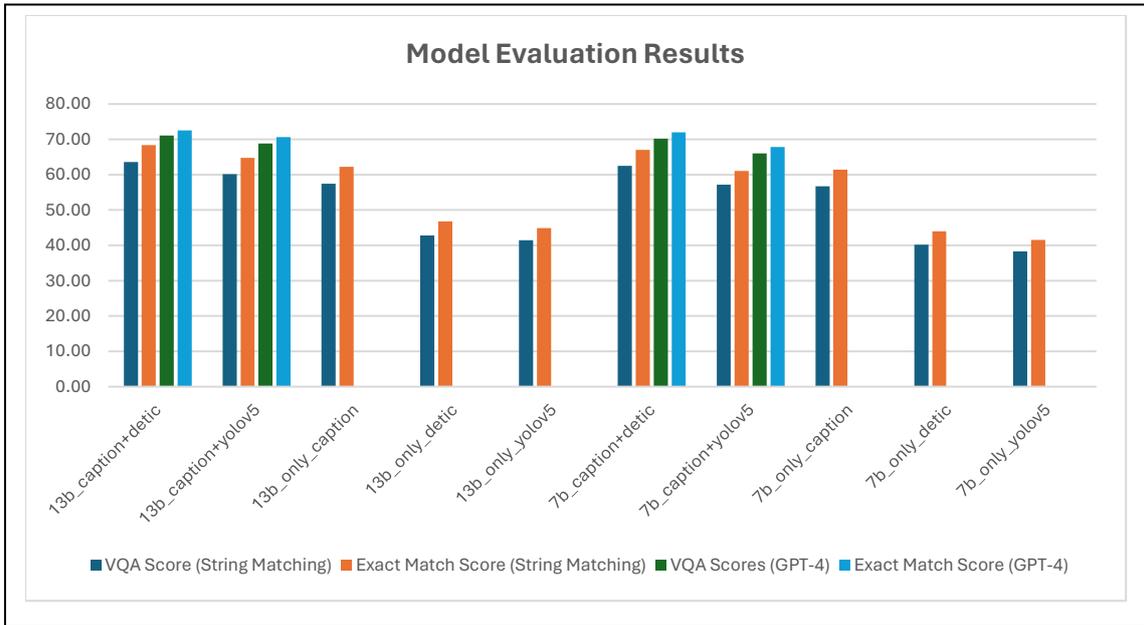


Figure 7-6 Ablation study results.

### C.3. Ablation Study Results per Question Category

Syntactic Evaluation (String Match)										
VQA Score										
Question Category	Fine-tuned LLaMA (13B)					Fine-tuned LLaMA (7B)				
	Caption+DETIC	Caption+YOLOv5	Only Caption	Only DETIC	Only YOLOv5	Caption+DETIC	Caption+YOLOv5	Only Caption	Only DETIC	Only YOLOv5
Brands, Companies and Products	50.98%	46.08%	44.12%	26.47%	21.57%	51.96%	47.06%	50.98%	27.45%	32.35%
Cooking and Food	64.96%	60.47%	57.26%	40.60%	33.33%	62.82%	53.21%	57.48%	42.09%	35.47%
Geography, History, Language and Culture	54.02%	52.87%	44.83%	33.33%	25.29%	52.87%	52.87%	50.57%	32.18%	27.59%
Objects, Material and Clothing	59.13%	53.17%	55.95%	38.89%	38.49%	57.14%	55.95%	55.56%	34.92%	31.35%
People and Everyday life	62.07%	57.47%	55.94%	37.93%	39.85%	58.62%	51.72%	44.06%	29.50%	32.57%
Plants and Animals	68.05%	66.86%	60.16%	47.34%	47.14%	68.84%	64.50%	61.14%	45.36%	42.01%
Science and Technology	50.98%	54.90%	60.78%	31.37%	39.22%	50.98%	47.06%	54.90%	35.29%	27.45%
Sports and Recreation	71.43%	71.15%	65.27%	61.06%	53.50%	73.39%	71.15%	70.03%	52.38%	52.38%
Vehicles and Transportation	57.86%	54.30%	52.62%	36.48%	41.72%	54.09%	54.72%	53.46%	34.59%	35.43%
Weather and Climate	52.56%	50.00%	47.44%	34.62%	30.77%	62.82%	53.85%	46.15%	47.44%	24.36%
Other	69.40%	62.57%	63.39%	46.17%	46.72%	66.94%	52.46%	56.28%	41.80%	44.26%
<b>Grand Total</b>	<b>63.57%</b>	<b>60.15%</b>	<b>57.49%</b>	<b>42.81%</b>	<b>41.42%</b>	<b>62.51%</b>	<b>57.19%</b>	<b>56.72%</b>	<b>40.19%</b>	<b>38.29%</b>
Exact Match Score										
Question Category	Fine-tuned LLaMA (13B)					Fine-tuned LLaMA (7B)				
	Caption+DETIC	Caption+YOLOv5	Only Caption	Only DETIC	Only YOLOv5	Caption+DETIC	Caption+YOLOv5	Only Caption	Only DETIC	Only YOLOv5
Brands, Companies and Products	55.88%	50.00%	47.06%	32.35%	26.47%	55.88%	52.94%	55.88%	29.41%	35.29%
Cooking and Food	69.87%	66.03%	62.18%	44.87%	37.82%	67.95%	57.05%	62.82%	46.15%	38.46%
Geography, History, Language and Culture	58.62%	55.17%	48.28%	34.48%	27.59%	58.62%	58.62%	55.17%	34.48%	31.03%
Objects, Material and Clothing	64.29%	57.14%	59.52%	41.67%	41.67%	60.71%	59.52%	58.33%	36.90%	34.52%
People and Everyday life	65.52%	60.92%	59.77%	41.38%	42.53%	62.07%	52.87%	47.13%	32.18%	34.48%
Plants and Animals	72.19%	70.41%	64.50%	50.89%	49.70%	73.37%	68.64%	65.68%	49.11%	44.97%
Science and Technology	58.82%	64.71%	70.59%	35.29%	47.06%	58.82%	52.94%	58.82%	41.18%	29.41%
Sports and Recreation	76.47%	76.47%	71.43%	67.23%	56.30%	78.15%	74.79%	74.79%	57.14%	56.30%
Vehicles and Transportation	62.26%	58.49%	57.23%	40.25%	45.28%	57.86%	57.86%	57.86%	38.36%	38.36%
Weather and Climate	57.69%	53.85%	50.00%	38.46%	34.62%	65.38%	57.69%	50.00%	53.85%	26.92%
Other	75.41%	68.85%	69.67%	50.00%	50.82%	72.95%	58.20%	63.11%	46.72%	49.18%
<b>Grand Total</b>	<b>68.36%</b>	<b>64.77%</b>	<b>62.28%</b>	<b>46.81%</b>	<b>44.91%</b>	<b>67.07%</b>	<b>61.08%</b>	<b>61.38%</b>	<b>44.01%</b>	<b>41.52%</b>

Table 7-3 Syntactic evaluation results for the ablation study per question category.

Semantic Evaluation (GPT-4)								
Question Category	VQA Score				Exact Match Score			
	Fine-tuned LLaMA (13B)		Fine-tuned LLaMA (7B)		Fine-tuned LLaMA (13B)		Fine-tuned LLaMA (7B)	
	Caption+DETIC	Caption+YOLOv5	Caption+DETIC	Caption+YOLOv5	Caption+DETIC	Caption+YOLOv5	Caption+DETIC	Caption+YOLOv5
Brands, Companies and Products	62.75%	53.92%	56.86%	51.96%	67.65%	55.88%	58.82%	52.94%
Cooking and Food	74.58%	69.87%	70.73%	62.40%	76.92%	71.79%	72.44%	64.74%
Geography, History, Language and Culture	59.77%	58.62%	66.67%	65.52%	62.07%	62.07%	68.97%	68.97%
Objects, Material and Clothing	67.46%	63.89%	65.08%	65.48%	67.86%	65.48%	66.67%	66.67%
People and Everyday life	68.97%	67.82%	67.43%	62.07%	70.11%	70.11%	67.82%	63.22%
Plants and Animals	74.36%	75.15%	74.95%	71.20%	75.15%	76.33%	76.33%	72.19%
Science and Technology	72.55%	72.55%	68.63%	66.67%	76.47%	76.47%	70.59%	70.59%
Sports and Recreation	76.19%	76.47%	79.83%	77.03%	77.31%	78.15%	81.51%	78.99%
Vehicles and Transportation	65.62%	63.73%	63.32%	65.20%	67.30%	66.04%	66.67%	67.92%
Weather and Climate	60.26%	61.54%	69.23%	66.67%	61.54%	61.54%	69.23%	69.23%
Other	75.41%	69.95%	72.95%	60.38%	76.23%	71.31%	74.59%	62.30%
<b>Grand Total</b>	<b>71.09%</b>	<b>68.86%</b>	<b>70.19%</b>	<b>65.97%</b>	<b>72.55%</b>	<b>70.66%</b>	<b>71.96%</b>	<b>67.86%</b>

Table 7-4 Semantic evaluation results for the ablation study per question category

### C.4. Additional Evaluation Samples

Figure 7-7 shows additional samples along with model answers including the ablation configurations.



Figure 7-7 Additional evaluation samples.

## D. LLaMA-2 Liscense by Meta

4/18/24, 5:34 PM

Get started with Llama 2 - Mohammed Bin Ali Alhaj - Outlook

Get started with Llama 2

Meta AI <noreply@email.meta.com>

Sun 11/5/2023 5:50 PM

To:mohas85@gmail.com <mohas85@gmail.com>

Llama 2 commercial license

### You're all set to start building with Llama 2.

The models listed below are now available to you as a commercial license holder. By downloading a model, you are agreeing to the terms and conditions of

#### Model weights available:

- Llama-2-7b
- Llama-2-7b-chat
- Llama-2-13b
- Llama-2-13b-chat
- Llama-2-70b
- Llama-2-70b-chat

With each model download, you'll receive a copy of the [License](#) and [Acceptable Use Policy](#), and can find all other information on the model and code on [Git](#)

#### How to download the models:

1. Visit [the Llama repository](#) in GitHub and follow the instructions in the [README](#) to run the download.sh script.
2. When asked for your unique custom URL, please insert the following:



3. Select which model weights to download

The unique custom URL provided will remain valid for model downloads for 24 hours, and requests can be submitted multiple times. Now you're ready to start building with Llama 2.

#### Helpful tips:

Please read the instructions in the GitHub repo and use the provided code examples to understand how to best interact with the models. In particular, for the you can find additional information about how to responsibly deploy Llama models in our [Responsible Use Guide](#).

#### If you need to report issues:

If you or any Llama 2 user becomes aware of any violation of our license or acceptable use policies - or any bug or issues with Llama 2 that could lead to ar

1. Reporting issues with the model: [Llama GitHub](#)
2. Giving feedback about potentially problematic output generated by the model: [Llama output feedback](#)
3. Reporting bugs and security concerns: [Bug Bounty Program](#)
4. Reporting violations of the Acceptable Use Policy: [LlamaUseReport@meta.com](#)

[Subscribe](#) to get the latest updates on Llama and Meta AI.

Meta's GenAI Team

كما طلبت: mohas85@gmail.com إرسال هذه الرسالة إلى  
Meta Platforms, Inc., Attention: Community Support, 1 Meta Way, Menlo Park, CA 94025