

IndiFoodVQA: Advancing Visual Question Answering and Reasoning with a Knowledge-Infused Synthetic Data Generation Pipeline

Pulkit Agarwal[†], Settaluri Lakshmi Sravanthi[†]

Pushpak Bhattacharyya[◇]

[◇]Indian Institute of Technology Bombay, [†] Equal contribution
{pulkitagarwal, sravanthi, pb}@cse.iitb.ac.in

Abstract

Large Vision Language Models (VLMs) like GPT-4, LLaVA, and InstructBLIP exhibit extraordinary capabilities for both knowledge understanding and reasoning. However, the reasoning capabilities of such models on sophisticated problems that require external knowledge of a specific domain have not been assessed well, due to the unavailability of necessary datasets. In this work, we release a first-of-its-kind dataset called IndiFoodVQA with around 16.7k data samples, consisting of explicit knowledge-infused questions, answers, and reasons. We also release IndiFoodKG, a related Knowledge Graph (KG) with 79k triples. The data has been created with minimal human intervention via an automated pipeline based on InstructBlip and GPT-3.5. We also present a methodology to extract knowledge from the KG and use it to both answer and reason upon the questions. We employ different models to report baseline zero-shot and fine-tuned results. Fine-tuned VLMs on our data showed an improvement of $\sim 25\%$ over the corresponding base model, highlighting the fact that current VLMs need domain-specific fine-tuning to excel in specialized settings¹. Our findings reveal that (1) explicit knowledge infusion during question generation helps in making questions that have more grounded knowledge, and (2) proper knowledge retrieval can often lead to better-answering potential in such cases.

1 Introduction

Visual Question Answering (VQA) was initially introduced as a mechanism to compare the ability of machines to behave like a human (Malinowski and Fritz, 2014b). Since the advent of chatbots like ChatGPT that show a high degree of understanding, they have become a common interface for human-machine interaction, where humans frequently ask questions based on specific domains to solve various problems. For instance, a restaurant

chatbot should excel in food-related queries and images, while fashion chatbots should specialize in recognizing delivered clothing items within images. While humans are extremely efficient at answering questions involving a single domain both before and after undergoing proper training, the same cannot always be said about language models. To develop such models, substantial domain-specific data is essential.

The primary necessity here is to get datasets that enable VLMs to show capabilities to understand and reason based on both prevalent and external knowledge. There have been numerous works pertaining to the requirement of commonsense knowledge (Johnson et al., 2017; Shah et al., 2019; Schwenk et al., 2022; Gao et al., 2022) in VQA, most using day-to-day images from datasets such as MS-COCO (Lin et al., 2014) and knowledge entities from generic KGs like ConceptNet (Speer et al., 2017). Only recently has attention grown towards a higher degree of reasoning according to knowledge in a particular area of interest (Lu et al., 2022; Wang et al., 2023). However, a big subset of curated datasets have been made by crowdsourcing efforts, which albeit being of high quality, are not easy to scale. With most state-of-the-art (SOTA) LLMs trained on huge chunks of data, this can be a big bottleneck.

In this work, we present a framework that leverages domain-specific knowledge and the superior capabilities of LLMs in text generation to create a reasoning benchmark with minimal human effort. Our contributions are:

1. **IndiFoodKG**: A Knowledge Graph based on recipes, ingredients, nutrients, and other miscellaneous data about Indian food dishes.
2. **IndiFoodVQA**: A multiple-choice visual question answering and reasoning dataset, created with IndiFoodKG as the underlying KG.

¹Data and code are available at [IndiFoodVQA](#).

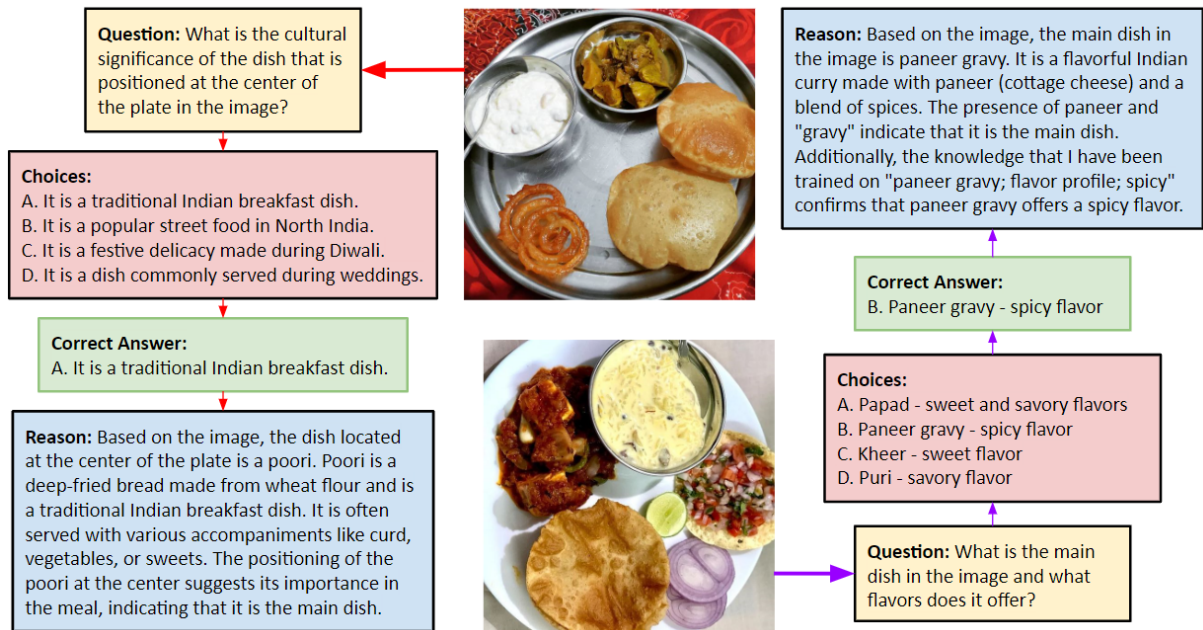


Figure 1: Examples from our dataset IndiFoodVQA, that require multiple reasoning steps. The second example shows a situation where the externally infused triples were used to reason on the generated question and answer.

3. A **knowledge-infused pipeline** to automatically generate questions from images; quality of the pipeline and effects of knowledge-infusion are discussed in Sections 3.3 and 4.5.
4. A **comprehensive evaluation** of IndiFoodVQA with various VLMs, performed for both zero-shot and fine-tuned models, with and without knowledge infusion.

The selection of the food domain, specifically Indian cuisine, is driven by its extraordinary diversity and daily significance. Present object detectors and vision encoders encounter challenges when tasked with identifying these food items, often confusing them with Western food dishes. This inherent bias also gets incorporated into the SOTA VLMs, which serves as additional motivation for choosing the niche domain of Indian food.

2 Related Work

VQA Dataset Generation. Approaches for Visual Question Generation (VQG) can be split into 2 buckets: human gold-standard datasets, and machine-generated datasets. We present different works from both approaches, along with their merits and shortcomings.

Human Annotated Datasets. The biggest drawback of this approach is quite evident - scalability issues, although it has still been the most

popular method (Mostafazadeh et al., 2016; Antol et al., 2015; Goyal et al., 2016; Krishna et al., 2017; Wang et al., 2017b; Marino et al., 2019). Another common idea here is to create human-annotated fixed question templates and simply replace certain words while making questions (Malinowski and Fritz, 2014a; Zhu et al., 2016; Yu et al., 2015). Although this could help increase the size of the dataset, it leads to a big decrease in the variability in questions and is not indicative of the real world where models should be able to answer a diverse set of questions.

Machine Generated Datasets. In Multitask iQAN Network (Li et al., 2018), the authors utilized the dual nature of VQA and VQG, by fusing the embeddings of the two modalities in an encoder-fusion-decoder module. Other important benchmarks in the visual reasoning space are CLEVR (Johnson et al., 2017), GQA (Hudson and Manning, 2019), and CRIC (Gao et al., 2022), created via automatic functional programs, which require reasoning over visual facts grounded in the image and facts found in external knowledge bases.

Multimodal Reasoning Benchmarks. The current benchmark in the space of reasoning is widely considered to be ScienceQA (Lu et al., 2022), consisting of multiple choice questions on various scientific topics along with corresponding answers, contexts, and explanations, created using heuristic

rules from open resources on science problems. A significant change in data generation methods was seen after LLaVA (Liu et al., 2023b) was released, which created multi-modal datasets using LLMs like GPT-3.5, with manually annotated captions and bounding boxes used to describe the image.

Knowledge-Based VQA. VQA based on external knowledge has been an important task, both to understand the capability that existing models have in terms of knowledge understanding and the limitations of using only inherent knowledge of the LLMs (Wu et al., 2016; Wang et al., 2017a; Narasimhan and Schwing, 2018; Cao et al., 2019; Gardères et al., 2020; Yu et al., 2020; Zhu et al., 2021; Shevchenko et al., 2021). Recent works have focused on external knowledge infusion, without changing the model weights. The KAPING framework (Baek et al., 2023) was developed to show that LLMs like T0 & GPT-3 injected with relevant knowledge triples through prompts attain superior zero-shot performance as compared to models using only internalized knowledge. Similarly, the Prophet framework (Shao et al., 2023) enabled GPT-3 to better comprehend the task of knowledge-based VQA by prompting with answer heuristics.

3 Knowledge Graph and Dataset

3.1 IndiFoodKG

We created a new KG called IndiFoodKG, with varied information about Indian food dishes. The KG has been compiled from three different sources:

- IndianFood101 (Prabhavalkar, 2020) - Information about 255 Indian dishes, their ingredients, place of origin, flavor profile, preparation time, and course of meal (2800 triples).
- CulinaryDB (Singh and Bagler, 2018) - Recipe to ingredient mapping of nearly 4k Indian food items (35k triples).
- Indian Food Composition Tables (Longvah et al., 2017) - Provides nutritional values for 528 key ingredients (42k triples).

Our curated knowledge graph has a total of 79, 934 unique triples, either accessing one of the 11 different relations or giving nutrient information about some ingredient. Each relation acts as a different specifier for a 1-hop triple. For example, the relation `has_ingredient` is a 1-hop triple between a dish and an ingredient. Details about the relations present in IndiFoodKG are given in Table 6.

3.2 IndiFoodVQA

We release IndiFoodVQA, a new benchmark in the field of knowledge-based VQA and reasoning. Each sample of IndiFoodVQA has 5 different parts: An image, a question based on the image, 4 possible answer choices, a correct answer out of the 4, and a reason for why the answer choice is correct.

Statistic	Number
Size of dataset	16, 716
Unique questions	13, 426
Question types	12
Number of images	414
Average question length	13.76
Average answer length	4.43
Average rationale length	59.23

Option A	Option B	Option C	Option D
5610	3929	3955	3222

Table 1: Important statistics for IndiFoodVQA - The second table represents the number of questions with the given option (A, B, C, or D) as the correct answer.

3.3 Quality Verification

To determine the extent of hallucination in the generated questions, we take 224 randomly chosen questions from the dataset, distributed equally across the different types of questions, and get them scored over 4 different aspects by human subjects. The task was divided among 20 people, with each data sample verified by 3 independent subjects to ensure inter-rater agreement. Every aspect is scored on a scale of 1 – 4, with a higher score indicating a better response. Specific instructions can be found in Appendix A.2. We obtained majority agreement (≥ 2 evaluators) across the 4 different questions asked to the subjects in 75% to 90% of the 224 data samples. The average scores are listed in Table 2. The human ratings are analyzed in detail in Appendix A.3.

Question relevance	Relevant choices	Correct answer	Correct reason
3.89	3.78	3.32	3.42

Table 2: Average scores on manual verification of 224 randomly chosen data samples on a scale of 1 – 4, considering only scores agreed upon by a majority.

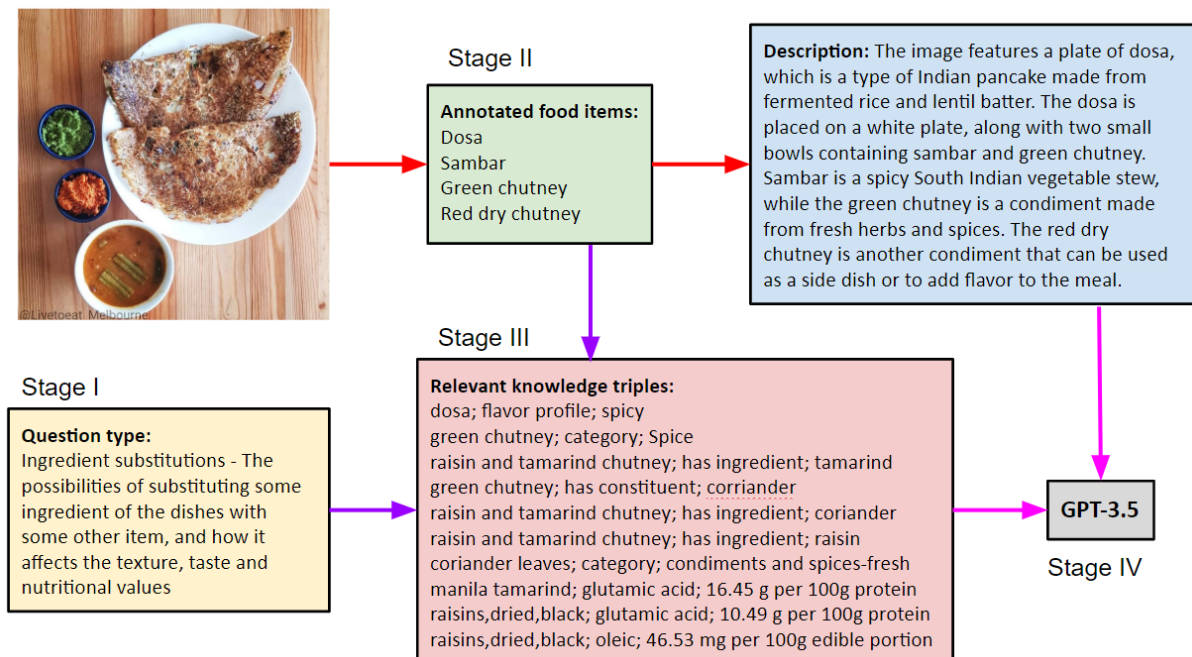


Figure 2: A 4-stage pipeline to automatically generate knowledge-based visual question reasoning dataset.

4 Question Generation Pipeline

4.1 Stage I: Question Type Templates

To ensure that questions generated by GPT-3.5 are related to our chosen domain, we create templates for different types of questions. The 12 templates were also created using ChatGPT and can be modified to fit any other domain. A detailed description of each type is given in Table 7. The prompt used to get the types can be found in Appendix C.1.

4.2 Stage II: Image Description

We first extract information from the images in natural language form. Unlike LLaVa (Liu et al., 2023b), which provides human-annotated captions and bounding boxes from MS-COCO (Lin et al., 2014) to GPT-3.5 for generating multi-modal data, we use machine-generated descriptions with human supervision. However, as we explain below, based on the domain being chosen this step can be performed without human intervention as well.

Human Annotation. We asked human annotators (details in Appendix A.1) to choose platter images from the IndianFood20 dataset (Goel et al., 2023) which have more than 3 items present in them. For each of the chosen 414 images, the annotators were asked to list down all the food dishes \mathcal{F} present in the image. This helped in guiding the description generation model to a relevant description of the image, which covers more visual

aspects. Note that this is a low-effort task, and is not an essential step in making the description.

Description Generation. We used the Instruct-Blip Vicuna-7B model (Dai et al., 2023) to create descriptions with the settings given in Appendix C.2. The model was prompted with the annotated food items \mathcal{F} and was asked to give a description \mathcal{D} of the color and relative location of those items. The description acts as an indicator of visual information in the image, which can not be inferred from knowing the food items alone.

4.3 Stage III: Knowledge Infusion

Before calling GPT-3.5 to generate questions, we also want to ensure that the questions will require knowledge from the KG to answer. We create a methodology for knowledge extraction to get triples \mathcal{T} from IndiFoodKG which are relevant to the image and the type of question. The triples are explicitly mentioned in the prompt given to GPT-3.5, without any verbalization, since past works (Moiseev et al., 2022; Baek et al., 2023) have shown that LLMs are capable of understanding these triples even if not in natural language form.

1-Hop Triples. We use embedding similarity to retrieve relevant triples, a technique that has been employed through graph embeddings in earlier works (Wang et al., 2014; Ma et al., 2019; Park et al., 2019; Nayyeri et al., 2023). Similar to KAP-

ING (Baek et al., 2023), the triples are linearly verbalized (subject, relation, and object joined via semi-colons as "s; r; o") to form elements of the corpora \mathcal{C} . The query sentence q is made using the annotated food items and the question type, again appended together with semi-colons. We use MPNet (Song et al., 2020) as our sentence embedding model for both q and the triples from \mathcal{C} , with cosine similarity as the metric for semantic distance.

To extract a more diverse range of triples, we retrieve separate triples from all the 3 knowledge sources mentioned in Section 3.1. The division of IndiFoodKG into the 3 knowledge bases can be done with the help of its relations, as described in Table 6. The top N triples which have the highest cosine similarity scores with the embedding of the query sentence, i.e. $\text{cos_sim}(q, \mathcal{C}, \text{top_}N)$, are the final retrieved triples $\mathcal{T}_{1\text{-hop}}$, where the hyperparameter N is chosen nearly in ratio with the size of each knowledge base. Thus, we take the top 5 triples from CulinaryDB (Singh and Bagler, 2018), the top 4 triples from the IFCT nutritional database (Longvah et al., 2017), and the top triple corresponding to IndianFood101 (Prabhavalkar, 2020), for a total of $N = 10$ triples.

2-Hop Triples. We utilize the structure of IndiFoodKG here, by which any 2-hop knowledge \mathcal{K}_2 about recipe-nutrient relation can be broken down into 1-hop relations about recipe-ingredient (\mathcal{K}_{r2i}) and ingredient-nutrient (\mathcal{K}_{i2n}) data. This idea is based on the inherent ability of LLMs like GPT-3.5 to combine two 1-hop triples and infer the corresponding 2-hop information, commonly enforced as chain-of-thought reasoning (Wei et al., 2022). Thus, instead of retrieving 2-hop knowledge, we simply find triples from IndiFoodKG with a common entity e (the ingredient).

To accomplish this, we first find all ingredients \mathcal{I}_{r2i} in IndiFoodKG which are from the CulinaryDB database (corresponding to recipe-ingredient relation). For each of these ingredients, we take its vector embedding (again with MPNet) as our query vector q_i . Similarly, we find all ingredients \mathcal{I}_{i2n} from the IFCT tables (corresponding to ingredient-nutrient data) and get their embeddings to create our corpus \mathcal{C}_i . The ingredient in the corpus with the highest cosine similarity score $\text{cos_sim}(q_i, \mathcal{C}_i, \text{top_}1)$ with a query ingredient is taken as the corresponding related entity \mathcal{I}_{re1} . To get our final top 10 triples, we again extract the top 1 and top 5 triples from IndianFood101 and

CulinaryDB respectively. Following this, for all the ingredients in the triples extracted so far, we find their related ingredient \mathcal{I}_{re1} . The nutrient information triples for these ingredients from the IFCT data are taken as our new corpus, and finally, we extract the top 4 triples only from these related triples. This ensures a higher degree of relation between the recipe-ingredient and ingredient-nutrient triples, and thus also gives a higher percentage of 2-hop information.

4.4 Stage IV: GPT-3.5 and Post-processing

We use the model gpt-3.5-turbo and provide it with the information sources from the previous 3 stages to influence its output - question type, image description, and the 2-hop extracted knowledge triples. The prompt and post-processing steps are given in Appendix C.3.

4.5 Impact of Knowledge Infusion

To comprehend the impact of KG infusion during question generation on the pipeline and its role in diversifying the question distribution, we quantify the number of questions influenced by the provided knowledge triples. For this, we first extract all noun words present in question or answer choices with the help of the spaCy library (Honnibal and Montani, 2017), and remove those words that were also present in the annotated food items. Finally, we check if any of these nouns are also present as a subject/object in the knowledge triples, or as one of the nutrients mentioned in the triples (for example words like "iron", "protein", "magnesium", etc.). 4050 questions in the dataset ($\sim 24\%$) were found to have added information from the knowledge graph, with the highest concentration in questions about health & nutritional aspects (649) and ingredients (608), and the least amount of knowledge infused into questions on the topics of cooking technique (91) and presentation & plating (56). This is in line with the kind of knowledge that IndiFoodKG has, showing that the knowledge infusion step was indeed successful in a large fraction of questions.

5 Experimental Setup

In this section, we describe the experimental setup used to establish the baselines. The dataset has been split into the train, validation, and test sets in a ratio of 70 : 10 : 20, thus consisting of 11, 709, 1661, and 3346 questions. The split into the test set has been done maintaining a roughly equal number

Model	Knowledge	Accuracy	Rouge-L	BLEU-1	BLEU-4	METEOR	Similarity
random	—	26.69	0.23	0.247	0.031	0.207	0.368
mplug-owl llama-7b (\mathcal{I})	No KG	34.13	0.302	0.33	0.095	0.325	0.824
	1-hop	32.22	0.291	0.313	0.09	0.325	0.807
	2-hop	32.82	0.289	0.31	0.089	0.325	0.806
	Original	33.32	0.29	0.31	0.091	0.34	0.811
open flamingo mpt-9b	No KG	25.46	0.093	0.034	0.0	0.06	0.517
	1-hop	31.05	0.078	0.023	0.0	0.047	0.497
	2-hop	28.06	0.076	0.022	0.0	0.045	0.488
	Original	29.23	0.075	0.023	0.0	0.045	0.483
instructblip flant5xxl- 11b (\mathcal{I})	No KG	52.06	0.172	0.022	0.006	0.089	0.715
	1-hop	50.57	0.217	0.044	0.014	0.123	0.738
	2-hop	50.75	0.212	0.035	0.012	0.118	0.732
	Original	54.15	0.217	0.033	0.013	0.121	0.747
llava llama2- 13b (\mathcal{I})	No KG	42.59	0.324	0.354	0.106	0.367	0.822
	1-hop	41.33	0.323	0.354	0.102	0.352	0.815
	2-hop	41.54	0.323	0.356	0.104	0.354	0.815
	Original	43.78	0.326	0.359	0.108	0.357	0.821

Table 3: Zero-shot evaluation on IndiFoodVQA. Accuracy is for the correct answer (in %). All other metrics are for the generated reason. Similarity refers to cosine similarity with the original reason using the Sentence-BERT model. The random model gives a random answer and a random reason from questions belonging to the same type in the train set, and \mathcal{I} under the model name stands for VLMs with an instruction-tuned base LLM. Knowledge refers to the type of triples presented to the models during inference, as explained in Section 5.1. No KG means inference without any external knowledge, 1-hop and 2-hop are for inference with the triples extracted by the corresponding method, and Original refers to inference with the triples given to GPT-3.5 during question generation. The bold values are the best accuracy scores by the 4 models and the best metric on reason generation across different models.

of questions of each question type. All results are reported for a single run of experiments.

5.1 Zero-Shot (ZS) Baselines

We benchmarked ZS baselines on VLMs ranging from sizes of 7B to 13B parameters: mplug-owl-llama-7b (Ye et al., 2023), openflamingo-mpt-9b (Awadalla et al., 2023), instructblip-flant5xxl-11b (Dai et al., 2023) and llava-llama2-13b (Liu et al., 2023b) as they have shown SOTA performance on various benchmarks. We also perform four types of evaluations on each model: no knowledge infusion, with extracted 1-hop knowledge triples, with extracted 2-hop knowledge triples, and when presented with the original knowledge triples given to GPT-3.5 during question generation.

- **Without knowledge infusion:** The model is given an image, a question, and 4 answer choices to predict and explain the answer.
- **With k -hop knowledge triples infusion:** In this method, the model again gets the image, question, and answer choices as input, along with knowledge triples up to k -hop ($k = 1, 2$)

added as a hint, with the aim of predicting the correct answer and a reason supporting the answer. The main idea is to exploit the fact that adding knowledge during LLM inference helps in improving understanding of the task (Liu et al., 2020; Zhang et al., 2022). To extract triples from IndiFoodKG corresponding to a given data sample, we use the same technique as given in Section 4.3 with a different query. The query sentence is made by extracting all noun chunks from the question, using spaCy (Honnibal and Montani, 2017) to extract these chunks. We ignore answer choices when finding the relevant triples since most of them will act as detractors, often leading to triples unrelated to the ones we desire.

- **With original GPT triples:** In this method, we evaluate the models if they are provided with the original triples given to GPT-3.5 (from Section 4.3). This is an ideal situation, where the exact same triples can be extracted.

For each model, we get the answer first, and the reason next after providing the generated answer

Paradigm	Knowledge	Accuracy	Rouge-L	BLEU-1	BLEU-4	METEOR	Similarity
No external triples	No KG	69.22	0.506	0.497	0.297	0.481	0.883
	1-hop	65.72	0.494	0.476	0.28	0.461	0.878
	2-hop	66.11	0.49	0.471	0.274	0.455	0.875
	Original	67.84	0.495	0.479	0.282	0.461	0.879
1-hop extracted triples	No KG	65.09	0.51	0.503	0.303	0.486	0.884
	1-hop	67.15	0.521	0.508	0.317	0.495	0.886
	Original	65.09	0.519	0.509	0.315	0.494	0.888
2-hop extracted triples	No KG	64.26	0.507	0.499	0.299	0.482	0.883
	2-hop	66.59	0.524	0.512	0.321	0.496	0.888
	Original	63.81	0.521	0.509	0.318	0.495	0.887

Table 4: Fine-tuned evaluation on IndiFoodVQA with llava-llama2-13b model. The model is fine-tuned under different paradigms as given in Section 5.2. The other details are the same as the ones explained in Table 3. For models fine-tuned along with 1/2-hop triples, we only perform inference with the corresponding triples.

to the model. The prompts and the technique used for all 4 models can be found in Appendix C.4. We also compare our scores with a random baseline, where we find all questions corresponding to the same question type from the train set, and choose a random answer and a random reason from this set,

5.2 Fine-Tuning (FT) Baselines

We benchmark FT baselines on llava-llama2-13b model fine-tuned on the train set. We perform three different types of fine-tuning setups, i.e. without any knowledge infusion, with 1-hop knowledge triples, and with 2-hop knowledge triples. When fine-tuning, both the answer and rationale are considered for the output. FT baselines are trained for 3 epochs on the existing instruction-tuned checkpoint of the model, with a learning rate of $2e-5$ and a global batch size of 128 (exact parameters are in Appendix D). The fine-tuned models are evaluated under the same 4 knowledge infusion paradigms as the ZS baselines.

5.3 Evaluation Metrics

For answer selection, we assess the top-1 accuracy, indicating the correctness of the chosen output among options A, B, C, and D. To evaluate the generated reasoning, we employ several metrics. These include the Rouge-L score (Lin, 2004), BLEU-1 and BLEU-4 (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005) scores, measured against the reasoning provided in the IndiFoodVQA dataset. Additionally, we include the sentence similarity score using Sentence-BERT (Reimers and Gurevych, 2019).

All experiments were performed on 2 NVIDIA A-100 GPUs. All models take 3 – 4 hours for

inference per task depending on the specific model being used, and 1 hour per epoch for training.

6 Results and Analysis

6.1 Baseline Scores

We report all results on the test set. The results for zero-shot evaluation and fine-tuned models are given in Table 3 and Table 4 respectively, for both answer selection and reason generation tasks. We also discuss a few other baselines (Yu et al., 2022; Liu et al., 2023a) in Appendix E. Several important points are evident from the results that we raise here, with discussions about improvements and future work deferred to Section 7:

- Across various models in zero-shot evaluation, a consistent observation is a slight dip in scores (approximately 2%) when incorporating extracted knowledge compared to inference without any knowledge. However, scores typically witness an improvement when the original triples are supplied to the model for answering. This underscores the potential for enhancing extraction methods that don’t solely rely on the question and involve image-level tokens and answer choices, potentially leading to improved performance.
- Instruction-tuned models such as mPLUG-Owl and LLaVA exhibit notable proficiency in reasoning. However, InstructBlip achieves the highest accuracy due to its training paradigm’s effective support for classification tasks (Wei et al., 2021), even though it struggles with verbalizing reasons. Conversely, models like OpenFlamingo, which possess a decoder-style


	<p>Question: Which ingredient in the image contains linoleic acid?</p> <p>Answer Choices: A) Bay Leaf B) Coriander C) Turmeric D) Onion</p>	<p>Correct Output: D) Onion</p> <p>Looking at the image, we can see that there are various food items present, including onion. According to the knowledge that I have been trained on provided, onion stalk contains linoleic acid. Therefore, the correct answer is D) Onion.</p>	<p>Base Model's Output: B) Coriander</p> <p>In the image, there is a dish with a variety of ingredients, including onions, coriander, and turmeric. The ingredient that contains linoleic acid is coriander. Coriander is a spice commonly used in Indian cuisine, and it is known to contain linoleic acid, which is an essential fatty acid. The presence of linoleic acid in coriander makes it a valuable ingredient in various dishes, including the one in the image.</p>	<p>Fine-Tuned Model's Output: D) Onion</p> <p>Based on the knowledge that I have been trained on onion stalk contains linoleic acid. Since the image shows onion as one of the ingredients, it can be inferred that onion in the image contains linoleic acid.</p>
---	---	--	--	---

Figure 3: Qualitative comparison of ground truth vs. generated zero-shot and fine-tuned answer and reason.

underlying LLM, demonstrate scores akin to near-random, reinforcing our inclination towards instruction-following models in the zero-shot context.

- Fine-tuning the base model on our dataset yields considerable enhancements in both generated answers and the quality of reasoning. This improvement signifies a better understanding of the domain when supplemented with relevant knowledge while training.
- Since we are testing the VLMs on a noisy machine-generated test set, we also create a *clean* test set (similar to Qasemi et al., 2023). For this, we used instances from the 224 verified samples which are from the test set and have a majority score of 4 (i.e. a majority of the raters claimed the sample is correct). There were a total of 98 such samples, and the best accuracy achieved by LLaVA zero-shot and fine-tuned models on this clean test set was 50.00% and 73.08% respectively, showing a similar improvement as the scores on the full test set.

6.2 Variation with Question Types

We also present the performance of different knowledge infusion techniques during inference with the 12 question types in Figure 4. In questions related to nutritional aspects, dietary restrictions, and ingredients, that saw the highest amount of knowledge infusion (Section 4.5), giving the correct knowledge is generally beneficial, highlighting the importance of extraction of appropriate triples. However, when considering open-ended questions about flavor profiles and presentation & plating, external unrelated knowledge can lead to a significant drop

in performance. This is mainly due to the tendency of these models to get influenced by the irrelevant triples, instead of being able to ignore them.

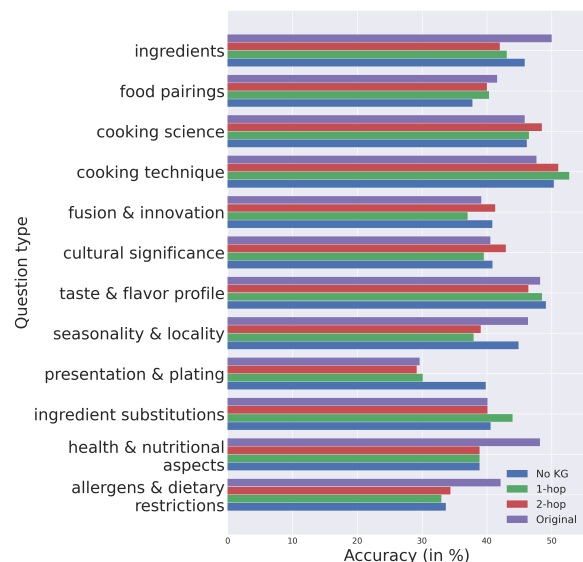


Figure 4: Accuracy scores (in %) for llava-llama2-13b model (zero-shot) across different question types.

6.3 Zero-Shot vs. Fine-Tuned

We qualitatively analyzed a few generated zero-shot and fine-tuned LLaVA outputs. A representative example, with both training and inference done using 2-hop triples, is given in Figure 3. The example serves as a clear indicator of how zero-shot modeling techniques are not enough when focusing on a specific domain. The base model gets affected by the distracting answer choices and incorrectly claims that coriander is present in the image. However, the fine-tuned checkpoint retrieves the correct information from the knowledge triples (which are the same for both the base and the fine-tuned model) and is able to output the right answer.

6.4 Object Detection Quality

We also analyze the extent to which model failures can be attributed to inaccuracies in detecting food items in the images. For the best-performing model (fine-tuned llava-llama2-13b), the output of 2972 samples from 3346 test set samples contains either the food items or the subject/object of the original triples that we provided to GPT-3.5, even though they were not provided in the question or answer choices. This establishes the fact that the VLM is generally able to detect the food items, implying that the low accuracies are majorly due to their inability to perform domain-specific reasoning based on external knowledge. As per our understanding, this is also influenced by the involvement of cues specific to the Indian cuisine in the question and answer choices, which help the model to focus along those directions.

7 Conclusion

We developed a novel domain-specific VQA generation pipeline using the existing large models and domain-specific knowledge from our curated KG IndiFoodKG. To the best of our knowledge, this is the first synthetic data generation pipeline that uses both external knowledge and the model’s internalized knowledge for creating VQA data. We have evaluated the performance of various baselines to establish the quality of the proposed dataset and showed how existing LLMs generally do not demonstrate good zero-shot performance when constrained to a domain. Our results showed a 15% improvement in accuracy with a fine-tuned LLaVA model over the best-performing zero-shot VLM.

Through this endeavor, our aim is to expedite multimodal research in fields where generating data at scale is a costly and labor-intensive task. Given the extensive training datasets used by contemporary LLMs, evaluating their effectiveness when incorporating external knowledge not present during training becomes increasingly critical. Assessing these models with knowledge pertaining to less-explored fields offers an optimal approach for such evaluation. Additionally, these datasets can serve as crucial benchmarks for detecting biases in SOTA VLMs. The architecture of our pipeline allows for seamless replacement of its components with elements from other domains, facilitating the creation of benchmarks and conducting studies in low-resource domains. Detailed insights into the generalizability of our model to diverse domains are dis-

cussed in Appendix F. Our research also prompts potential modifications in both retrieval and modeling techniques to enhance the off-the-shelf domain-relevant performance of versatile LLMs.

8 Limitations

One clear limitation of the IndiFoodVQA dataset and the knowledge-infused pipeline is the exclusive use of the English language, which limits its accessibility and usability for non-English speakers and in regions where English is not widely spoken, that can become important when restricting the environment to a specific domain. Another limitation is the requirement of OpenAI API access (as we have used GPT-3.5 as a major component of the data generation pipeline). However, this can be overcome by replacing GPT-3.5 with any openly available large foundational models like Llama 2 (Touvron et al., 2023) or Falcon-180b (Almazrouei et al., 2023).

We also note that the KG covers only a subset of the topics that are used for creating the questions. For example, there are very few knowledge triples on ‘cultural significance’ in IndiFoodKG (Table 6), so any questions that GPT-3.5 comes up with from that category are neither grounded in KG nor can be answered completely using the KG. This is not necessarily a drawback of the dataset, but it cannot be expected that models will improve dramatically simply with the infusion of our KG. To show large improvements, the pretrained knowledge of the model itself will need to be greatly expanded and that’s simply not the case with most open-source LLMs today. Alternatively, models need to get access to the relevant knowledge, so the source of external knowledge cannot be just the IndiFoodKG knowledge base. We further discuss this issue by using the generate-then-read method (Yu et al., 2022) in Appendix E.1. When generalizing to a different domain, this can be mitigated by choosing question categories that are highly grounded in the knowledge available.

Ethics Statement

This research was conducted in accordance with the ACL Ethics Policy. The ethical considerations during both the human annotation and verification process are discussed in Appendix A.

References

- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Maitha Alhamadi, Mazzotta Daniele, Daniel Hessel, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of language models: Towards open frontier models.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.
- Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. [Knowledge-augmented language model prompting for zero-shot knowledge graph question answering](#). In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106, Toronto, Canada. Association for Computational Linguistics.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Qingxing Cao, Bailin Li, Xiaodan Liang, and Liang Lin. 2019. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network. *arXiv preprint arXiv:1909.10128*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#).
- Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2022. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- François Gardères, Maryam Ziaefard, Baptiste Abeoos, and Freddy Lecue. 2020. [ConceptBert: Concept-aware representation for visual question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online. Association for Computational Linguistics.
- Mansi Goel, Shashank Dargar, Shounak Ghatak, Nidhi Verma, Pratik Chauhan, Anushka Gupta, Nikhila Vishnumolakala, Hareesh Amuru, Ekta Gambhir, Ronak Chhajed, et al. 2023. Dish detection in food platters: A framework for automated diet logging and nutrition management. *arXiv preprint arXiv:2305.07552*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. [Making the V in VQA matter: Elevating the role of image understanding in visual question answering](#). *CoRR*, abs/1612.00837.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.
- Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6116–6124.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. [Llava-1.6: Improved reasoning, ocr, and world knowledge](#).

- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. [Visual instruction tuning](#). In *NeurIPS*. Oral Presentation Project Page: <https://llava-vl.github.io/>.
- Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. [K-bert: Enabling language representation with knowledge graph](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2901–2908.
- Thingnganing Longvah, Rajendran Ananthan, K Bhaskar, and K Venkaiah. 2017. [Indian food Composition Tables](#).
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Minbo Ma, Fei Teng, Wen Zhong, and Zheng MA. 2019. [A sentence-rcnn embedding model for knowledge graph completion](#). In *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 484–490.
- Mateusz Malinowski and Mario Fritz. 2014a. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27.
- Mateusz Malinowski and Mario Fritz. 2014b. Towards a visual turing challenge. *arXiv preprint arXiv:1410.8027*.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.
- Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. [SKILL: Structured knowledge infusion for large language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588, Seattle, United States. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. [Generating natural questions about an image](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany. Association for Computational Linguistics.
- Medhini Narasimhan and Alexander G Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 451–468.
- Mojtaba Nayyeri, Zihao Wang, Mst. Mahfuja Akter, Mirza Mohtashim Alam, Md Rashad Al Hasan Rony, Jens Lehmann, and Steffen Staab. 2023. [Integrating knowledge graph embeddings and pre-trained language models in hypercomplex spaces](#). In *22nd International Semantic Web Conference (06/11/23 - 10/11/23)*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Junseok Park, Kwangmin Kim, Woochang Hwang, and Doheon Lee. 2019. [Concept embedding to measure semantic relatedness for biomedical information ontologies](#). *Journal of Biomedical Informatics*, 94:103182.
- Neha Prabhavalkar. 2020. [Indian food 101](#).
- Ehsan Qasemi, Amani R Maina-Kilaas, Devadutta Dash, Khalid Alsaggaf, and Muhao Chen. 2023. Preconditioned visual language inference with weak supervision. *arXiv preprint arXiv:2306.01753*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer.
- Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.
- Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.
- Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. 2021. [Reasoning over vision and language: Exploring the benefits of supplemental knowledge](#). In *Proceedings of the Third Workshop on Beyond Vision and LAnguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 1–18, Kyiv, Ukraine. Association for Computational Linguistics.
- Navjot Singh and Ganesh Bagler. 2018. [Data-driven investigations of culinary patterns in traditional recipes across the world](#).

- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. **Llama 2: Open foundation and fine-tuned chat models**. *CoRR*, abs/2307.09288.
- Min Wang, Ata Mahjoubfar, and Anupama Joshi. 2023. Fashionvqa: A domain-specific visual question answering system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3513–3518.
- Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. 2017a. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI’17*, page 1290–1296. AAAI Press.
- Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017b. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.
- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. **Knowledge graph and text jointly embedding**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1591–1601, Doha, Qatar. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4622–4630.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Jing Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition*, 108:107563.
- Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.
- Taolin Zhang, Chengyu Wang, Nan Hu, Minghui Qiu, Chengguang Tang, Xiaofeng He, and Jun Huang. 2022. Dkplm: decomposable knowledge-enhanced pre-trained language model for natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11703–11711.
- Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.
- Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2021. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI’20*.

A Human Annotation & Evaluation

A.1 Annotation

To choose our initial set of images, as well as to get annotated food items present in those images, we procured the help of 2 annotators with sufficient knowledge of Indian food dishes. From the Indian-Food20 dataset (Goel et al., 2023), the annotators were assigned 10 dish classes each and asked to select 21 images from each dish class. The only constraint during image selection was to search for images with at least 3 food items present in them. Then each image was annotated by the corresponding annotator, and the annotated food dishes were verified by the other annotator. We removed any images where there was a disagreement between the 2 annotators. The final image set consisted of 414 images. The annotations were performed independently, and each annotator received 0.5 USD for each sample they annotated.

A.2 Manual Verification

We consulted 20 human subjects for the verification of a random subset of our data, with all subjects highly qualified, either having completed or currently pursuing a bachelor’s degree in their final or pre-final year. The evaluators were asked 4 different questions about the dataset, as shown in Figure 5, and were supposed to give a score from 1 to 4 for the same. Each participant was adequately compensated for the task, being paid up to 0.15 USD for each evaluated question. During the final average score calculations, we swapped the scores of 1 and 2, to give more weightage to the confidence of the participants in their scores.

Each question was scored independently by 3 different evaluators, without access to the scores provided by each other, and majority agreement was considered before determining the scores. Out of the 224 samples chosen for manual verification, the 4 questions had an inter-rater agreement for 198, 198, 174, and 166 data samples respectively. For the final scores, as provided in Table 2, we found the average over these majority-agreed samples.

A.3 Analysis of Human Ratings

We performed a more detailed error analysis to understand the reason why some samples were provided with low scores by the human evaluators. This is presented in Table 5, for the 224 human-rated samples.

Error type	% of samples
Hallucination due to incorrect visual features in description	8.57
Hallucination by GPT-3.5	18.09
Presence of closely related food items or answer choices	3.81
Presence of a question with a highly subjective answer	15.24

Table 5: Analysis of human-rated samples.

Here, we have classified the samples on which a majority of human raters gave a score lower than 3 for one of the questions asked to them. The remaining 54.29% of evaluated samples received a high majority score across all the four questions asked to the evaluators. We notice that the last two reasons for low ratings in Table 5 are highly dependent on the human subject, which means that only around 26% of the samples had a low rating in some aspect due to hallucination by the pipeline.

The inter-rater agreement during the manual evaluation was low for metrics like ‘correct answer’ and ‘correct reason’ (Table 2). We noticed that while calculating the agreement scores for these aspects, we did not filter the samples that received low scores in Q1 and Q2 (Figure 5). Therefore the error gets accumulated for the scores of Q3 and Q4. If we only consider those ratings that correspond to a correct question and correct choices (i.e. 4 in the first two questions – there are 204 such instances out of the 224 manually verified samples), then the scores for ‘correct answer’ and ‘correct reason’ become 3.55 and 3.53 respectively.

B KG and Dataset

B.1 IndiFoodKG Relations

We present all the relations from IndiFoodKG in Table 6, along with their source knowledge base, and the number of triples corresponding to each relation.

B.2 Question Types

We list down all 12 types of questions that have been considered in the dataset in Table 7.

The short description (keywords) are used when making the query sentence for KG triple extraction as described in Section 4.3. The long description is used in the prompt for GPT-3.5 given in Appendix C.3.

Not sure 1 | Incorrect 2 | Not completely correct 3 | Completely correct 4

Please answer these questions:

Q1: Does the question make sense with the image? **A1:**

Q2: Is the answer in one of the given choices? **A2:**

Q3: Is the answer correct for the given question? **A3:**

Q4: Is the reason behind the given answer correct? **A4:**

Figure 5: Questions asked to the human subjects for manual verification of IndiFoodVQA.

Relation	Meaning	Source	# Triples
preparation_time	Time needed to prepare a dish	IndianFood101	225
cooking_time	Time needed to cook the dish	IndianFood101	227
flavor_profile	Spicy, sweet, sour, etc.	IndianFood101	226
found_in_state	Indian state where dish is found	IndianFood101	231
course_of_meal	Main course, snack, dessert, etc.	IndianFood101	255
type_of_diet	Vegetarian or non-vegetarian	IndianFood101	255
from_region	Region of India where dish is found	IndianFood101	242
has_ingredient	Ingredients present in a recipe	CulinaryDB	34,020
category	Ingredient types (poultry, seeds, etc.)	CulinaryDB	1530
synonym	Other names used for an ingredient	CulinaryDB	600
has_constituent	Constituent ingredients	CulinaryDB	448
Others	Nutrient information of ingredients	IFCT	41,674

Table 6: Relations present in IndiFoodKG.

C Prompts

C.1 Question Type Templates

We prompted ChatGPT to get the different question types along with a detailed description of each (a total of 12 types have been considered).

The task is to design templates for different question types to be present in Indian food VQA. Suggest some templates for different question types. Also give descriptions for each template.

We generated a few template types for the questions using ChatGPT, which provided us with 18 such unique question types over 3 runs. 12 were chosen as relevant ones based on advice from domain experts as well as to avoid too much intersection between questions of different types. Other generated templates were identification (not reasoning-based, more focused towards object detection), spice level (discarded because it was cov-

ered through flavor profile), historical evolution (discarded by nutritionist), sustainability (discarded by nutritionist), regional variations (discarded by nutritionist), and culinary influences (similar to fusion and innovation).

C.2 Description Generation

The description for the image is generated using InstructBlip Vicuna-7B model, with the following prompt and settings:

The following food items are present in this image: {annotated food items}. Describe the color and relative location of each food item in detail.

- num_beams = 3
- max_length = 300
- min_length = 1
- top_p = 0.9
- repetition_penalty = 3.0

Question type	Keywords	Detailed description
ingredients	ingredients, overall flavor and aroma of the dish	what are the key ingredients and their roles in the food items, and how do they contribute to the overall flavor and aroma of the dish
cooking technique	cooking technique, impact on preparation time, color, texture and flavor	how does the cooking technique differ from other similar dishes, and how does it impact preparation time, color, texture and flavor of the dishes
cultural significance	cultural significance, Indian festivals, seasonal produce	what is the cultural significance of the dishes in Indian festivals, and how does it reflect the celebration of seasonal produce
taste and flavor profile	taste and flavor profile, balance of sweet, savory, and spicy flavors	how do these items create a balance of sweet, savory, and spicy flavors, and how does this diversity enhance the dining experience
health and nutritional aspects	health and nutritional benefits, protein, fiber, nutrient and mineral content	how do the nutritional benefits compare with other similar dishes, highlighting the protein, fiber and other nutrient and mineral content in each food item
seasonality and locality	seasonality and locality, regional spices	what kind of regional spices and ingredients are generally used, and how it connects to the local produce of the states in which these dishes are generally consumed
ingredient substitutions	ingredient substitutions, similarities	the possibilities of substituting some ingredient of the dishes with some other item, and how it affects the texture, taste and nutritional values
presentation and plating	presentation, plating and garnishing	the importance of garnishing and presentation in the dishes, and how it impacts the overall dining experience
fusion and innovation	fusion and blending with other cuisines and innovation	how the given food items can be combined with other cuisines, and how the blending of ingredients from different cultures can create a unique culinary experience
cooking science	cooking science, scientific processes	what scientific processes might be involved in making these food items, and how it affects the texture and taste of the final product
allergens and dietary restrictions	allergens and dietary restrictions, alternative ingredients or preparation methods to make it allergen-free	what is the allergen content in the food items, and alternative ingredients or preparation methods to make it allergen-free
food pairings	traditional pairing of other complementary food dishes	traditional pairing of other food dishes with the food items shown, and how these complement with each other

Table 7: The 12 different question types.

- length_penalty = 1.2
- temperature = 1

C.3 Question Generation using GPT-3.5

The prompt given to GPT-3.5 for generating questions is inspired by the prompt used in (Liu et al., 2023b), modified according to the domain of food items, and keeping in mind our explicit knowledge infusion step.

You are an Indian food specialist AI visual assistant, and you are seeing a single image. What you see are provided with some sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Description: {image description}

Use the following facts when generating the questions, given in the form of triples:
{KG triples}

Give an output with 4 parts, with each part separated by 2 blank lines: a question (name it Question, and give the question in the next line), 4 possible answer choices (name it Answer Choices, with choices A, B, C and D in separate lines), the correct answer to that question (name it Correct Answer, out of A, B, C and D), and a reason for that answer (name it Reason, limited to 1 paragraph). Ask diverse questions and give corresponding answers. Give me 5 such questions as output. Only include questions that have definite answers:

- (1) one can see the content in the image that the question asks about and can answer confidently
- (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

The question should be about {question type} of the food items in the image. This includes details about {detailed information about question type}. The question should involve complex ideas like relative positions of the objects, the shapes and colors of the objects, and so on. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Nowhere should it be mentioned that a description or some external knowledge has been provided. Act like you can see the image, and create complex questions requiring multiple steps of reasoning.

The knowledge triples do not describe the image. If any of the given knowledge triples are used to generate the question, then do not mention the entities given in the knowledge triple in the Question or Answer Choices. Ensure that in the case that any knowledge triple is used, the question is not answerable without using this external knowledge. The knowledge used to generate the question can only be mentioned in the Reason field.

Also, create questions about both the main dish and the side dish. Try to include the relative position between the items as a part of the question. But keep the main question about {question type} of the food items. Do not bold anything (keep everything in normal font), and do not number the questions. The question and each answer choice should be in a new line. Make sure the questions involve reasoning to answer. The output should contain 5 such diverse questions (5 questions with given format). Do not mention the word "knowledge" or "triples" or "description" anywhere. Don't include any numbers anywhere.

To maximize the diversity of questions as well as the utilization of the number of questions per prompt, 5 questions are requested for each output. We also experimented with 3 different temperature settings - 0.2, 0.4, and 0.7. Based on qualitative analysis of the generated questions, we chose the final temperature as 0.4, due to its ability to give a variety of questions.

After getting the output, we process the questions and replace words like "description", "knowledge triples", and "mentioned" with "image", "knowledge that I have been trained on", and "shown". We also remove any questions whose correct answer has been given as "Not answerable by the image", instead of as one of the 4 answer choices.

C.4 Zero-Shot Models

For all models, we use a 2-step answering methodology. First, the model is prompted with the question and the answer choices (along with any triples to be provided). We consider the output as the generated answer and again prompt the model to create a rationale behind this answer. All models are run with a limit on the maximum number of new tokens to 256 during rationale

Model	Answer prompt	Rationale prompt
mplug-owl-llama-7b	Answer prompt #1	Rationale prompt #1
openflamingo-mpt-9b	Answer prompt #2	Rationale prompt #2
instructblip-flan5xxl-11b	Answer prompt #2	Rationale prompt #1
llava-llama2-13b	Answer prompt #2	Rationale prompt #2

Table 8: Prompts used for inference by different models during zero-shot evaluation as given in Appendix C.4.

generation. We use the following prompts, which are shared across the 4 models.

Answer prompt #1:

Below are facts in the form of triples that might be meaningful to answer the question -
{extracted triples}

{question}
{answer choices}

Choose one correct answer for the question out of the 4 answer choices above.
Is the answer A, B, C or D?

The model’s output starts with "The answer is _" where _ is chosen out of A, B, C, and D. Any output not of this form is taken as incorrect.

Answer prompt #2:

Below are facts in the form of triples that might be meaningful to answer the question -
{extracted triples}

Focus less on the given triples.

{question}
{answer choices}

Given the image, choose one answer out of A,B,C,D. Answer:

The first letter is taken as the correct answer (will be one of A, B, C, or D).

Rationale prompt #1:

Below are facts in the form of triples that might be meaningful to answer the question -
{extracted triples}

{question}

The correct answer is {generated answer }.

Why? Explain in a short paragraph.

We removed any unfinished sentences from the rationale and extracted only the first paragraph as the generated reason, to keep the output concise (similar to the ground truth reason generated by

GPT-3.5).

Rationale prompt #2:

Below are facts in the form of triples that might be meaningful to answer the question -
{extracted triples}

Focus less on the given triples.

{question}
{answer choices}

The correct answer is {generated answer }.

Why? Explain with a detailed reason behind the given answer. Do not repeat any words from the given answer. Reason:

Prompts used by different models. Table 8 shows the different prompts used by each of the 4 models during the 2-step prompting process.

D Fine-tuned models

When fine-tuning the LLaVA model, we use a single prompt for both answering and reasoning. The prompt used is the same as Answer prompt #1 in Appendix C.4. The training is done to get the answer and the reason directly in separate lines, so we don’t need to use a 2-step prompt. Below are the hyperparameters used for fine-tuning the model:

- bf16 = True
- number_of_training_epochs = 3
- per_device_eval_batch_size = 4
- per_device_train_batch_size = 8
- gradient_accumulation_steps = 8
- learning_rate = 2e-5
- weight_decay = 0.
- warmup_ratio = 0.03
- lr_scheduler_type = "cosine"

E Other Baselines

A few days prior to the submission of this paper, two additional versions of the LLaVA model were

Model	Accuracy	Rouge-L	BLEU-1	BLEU-4	METEOR	Similarity
LLaVA (zero-shot) without any KG	42.59	0.324	0.354	0.106	0.367	0.822
LLaVA fine-tuned without any KG	69.22	0.506	0.497	0.297	0.481	0.883
LLaVA (zero-shot) with GPT-3.5 knowledge	59.379	0.426	0.447	0.212	0.432	0.862
LLaVA fine-tuned on GPT-3.5 knowledge	70.233	0.510	0.500	0.302	0.485	0.886

Table 9: Comparative performance analysis of LLaVA models employing various approaches. The comparison is done across both zero-shot and fine-tuned settings, when not using any knowledge vs. when the knowledge generated by GPT-3.5 is used (Appendix E.1). The other details are the same as the ones explained in Table 3.

Question Type	Accuracy (Fine-tuned w/o KG)	Accuracy (Fine-tuned <i>genread</i>)
allergens and dietary restrictions	60.70	60.0
cooking science	79.60	77.93
cooking technique	84.80	84.12
cultural significance	73.65	73.99
food pairings	55.87	62.86
fusion and innovation	67.23	68.94
health and nutritional aspects	65.45	67.44
ingredient substitutions	71.50	63.77
ingredients	71.18	67.71
presentation and plating	58.8	67.71
seasonality and locality	63.14	67.15
taste and flavor profile	75.45	77.54

Table 10: Accuracy scores (in %) for *genread* baseline across different types of questions.

introduced: LLaVA-1.6 with 34B parameters (Liu et al., 2024) and LLaVA-RLHF (Sun et al., 2023). Given the proximity of their release to our paper submission, we had insufficient time to conduct experiments with these models on our dataset. It remains intriguing to examine their performance in addressing the task at hand.

E.1 Generate-then-Read Baseline

We evaluated our dataset using the generate-then-read method (Yu et al., 2022), with GPT-3.5 as the generator, and our best LLaVA model (i.e. the fine-tuned model) as the reader. We first generated image descriptions using the fine-tuned LLaVA model, which we provided to GPT-3.5 along with the question and answer choices. We then prompted the model to generate relevant background knowledge that would be useful to answer

the question. We performed zero-shot inference with this knowledge added to the prompt on the fine-tuned LLaVA model. We also fine-tuned the base LLaVA model along with this knowledge. The results are reported in Table 9.

We observe that the generate-then-read (*genread*) technique is able to outperform the best score using knowledge from IndiFoodKG, when the LLaVA model is fine-tuned along with the generated knowledge. However, a more detailed analysis of the change in accuracies across different question categories (Table 10) shows that an increase in accuracy is generally shown in question types with highly subjective questions, such as presentation and plating. This is a result of the infusion of external knowledge (from IndiFoodKG) in the questions, as well as the fact that pre-trained LLMs

do not have the necessary knowledge to answer such domain-specific questions.

E.2 LLaVA-1.5

The newly introduced LLaVA-1.5 model (Liu et al., 2023a) is purported to demonstrate SOTA performance across 11 benchmarks despite being trained on a relatively smaller dataset. Our evaluation involved testing the model’s performance on IndiFoodVQA, and comparing it with the performance by LLaVA-2. The results are provided in Table 11.

Triples	LLaVA-2	LLaVA-1.5
No KG	42.59	33.21
1-hop	41.33	32.45
2-hop	41.54	32.00
Original	43.78	33.46

Table 11: Accuracy (in %) of zero-shot evaluation using LLaVA-1.5 and LLaVA-2. The other details are the same as the ones explained in Table 3.

We observe that, contrary to the claim made by the authors for other benchmarks, LLaVA-1.5 is not able to achieve similar zero-resource performance as LLaVA-2 on the given dataset. This discrepancy can be attributed to the presence of questions that necessitate comprehensive inherent knowledge of LLMs for accurate answering – specifically, questions for which IndiFoodKG lacks pertinent information. Nevertheless, the trends shown in different types of knowledge infusion remain the same, indicating that effective knowledge retrieval can still be beneficial.

F Generalizability of the Pipeline

Because of the way our pipeline has been structured, it has the potential to replace IndiFoodKG with some other KG, while maintaining the quality of the pipeline. Our work shows one possible application of the pipeline, along with experiments on some models to understand its intricacies. We also note that our pipeline can be extended to other domains, with certain changes in the approach, that we describe below:

1. Question types - Based on the domain, relevant types will be required. This may be done by human domain experts or using some machine generation followed by manual verification (which is what we did).

2. Image description - This step may require human intervention based on the domain. In our example, we used human annotators to find the food items, so as to shift the description along that direction. For a different domain, either a similar approach can be used (i.e. giving some relevant entities from the image to a description-generating model), or one can get descriptions from human domain experts.
3. Knowledge infusion - This step requires the presence of a KG pertaining to that domain and a method to extract relevant triples from the image description and question types.
4. Generation of data samples - This stage can be easily done for any other domain using the data generated in the previous stages, with a similar prompt as used for IndiFoodVQA (Appendix C.3).

Currently, we are providing 2-hop knowledge from the KG while generating the questions to ensure that the model requires more than one step of reasoning during inference. This can be adapted or extended to other domains based on the way knowledge is extracted from the relevant KG. Our prompt and description also help make questions that involve details about relative positions and colors/shapes of the food items, requiring various logical reasoning steps to answer. Similar techniques can be used in other domains, by having specific logical information in the description and prompting GPT-3.5 towards using that information during question generation.