The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

KAM-CoT: Knowledge Augmented Multimodal Chain-of-Thoughts Reasoning

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda,
Rituraj Singh, Godawari Sudhakar Rao

Samsung R&D Institute India - Bangalore
{d.mondal, suraj.modi, subha.darshi, rituraj.s, g.sudhakar}@samsung.com

Abstract

Large Language Models (LLMs) have demonstrated impressive performance in natural language processing tasks by leveraging chain of thought (CoT) that enables step-by-step thinking. Extending LLMs with multimodal capabilities is the recent interest, but incurs computational cost and requires substantial hardware resources. To address these challenges, we propose KAM-CoT a framework that integrates CoT reasoning, Knowledge Graphs (KGs), and multiple modalities for a comprehensive understanding of multimodal tasks. KAM-CoT adopts a two-stage training process with KG grounding to generate effective rationales and answers. By incorporating external knowledge from KGs during reasoning, the model gains a deeper contextual understanding reducing hallucinations and enhancing the quality of answers. This knowledge-augmented CoT reasoning empowers the model to handle questions requiring external context, providing more informed answers. Experimental findings show KAM-CoT outperforms the state-of-the-art methods. On the ScienceQA dataset, we achieve an average accuracy of 93.87%, surpassing GPT-3.5 (75.17%) by 18% and GPT-4 (83.99%) by 10%. Remarkably, KAM-CoT achieves these results with only 280M trainable parameters at a time, demonstrating its cost-efficiency and effectiveness.

Introduction

Large Language Models (LLMs), particularly GPT-3 (Kojima et al. 2022a), ChatGPT (OpenAI 2022) and recently LLaMA, LLaMA2 (Touvron et al. 2023a,b) have demonstrated exceptional performance in natural language processing tasks. Additionally, incorporation of chain of thought (CoT) method in LLMs has revolutionized the way machines approach reasoning intensive tasks (Zhou et al. 2023). CoT refers to the ability of LLMs to think and reason in a step-by-step manner, mirroring the human cognitive processes (Wei et al. 2022b). Traditional language models (LMs) generate responses without explicit intermediate steps, which may lead to sub-optimal answers, especially in complex reasoning scenarios. CoT addresses the limitations by enabling language models to reason by introducing intermediate steps, thereby enhancing the model's problem-solving capabilities.

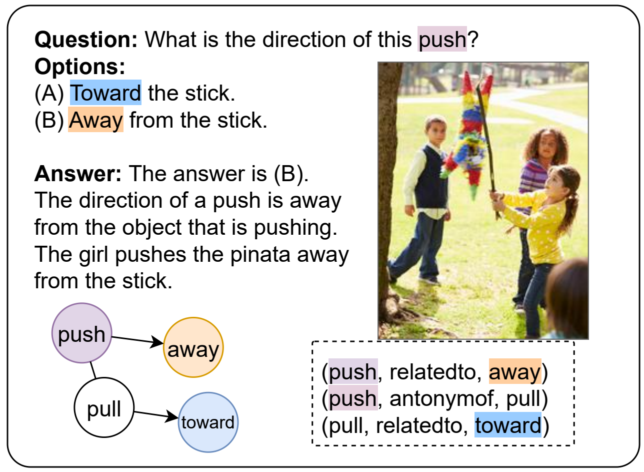

Figure 1: An example from ScienceQA dataset (Lu et al. 2022) showing how graphs can aid in multi-modal QA.

Recently, there is a surge to extend LLMs with multimodal capabilities. The fusion of visual and textual information has led to significant advancements in vision-and-language tasks, like visual question answering (VQA), image captioning, and image-text retrieval, and has opened up potential for transformative progress. Authors Liu et al. (2023a); Gao et al. (2023); Lu et al. (2023a) recognize and advocate the value of amalgamating visual and linguistic modalities. However, the behemoth scale of these models necessitates substantial computational resources, particularly in terms of hardware infrastructure. Zhang et al. (2023c) proposes fine-tuning smaller models to adapt to multimodality and elicit CoT capabilities. Nevertheless, such an approach tends to result in hallucinations, where the model generates plausible, but incorrect reasoning and answers. One possible solution is to integrate Knowledge Graphs (KGs) for enhancing model comprehension.

KGs serve as valuable structured knowledge sources, capturing information from various domains. For CoT reasoning, KGs can supplement step-by-step reasoning. By incorporating information from KGs, language models can reason more coherently, and leverage contextual relationships between entities and attributes. Consider the question in Figure 1. The knowledge about the direction of push is pivotal to answer the question. The KG triples (shown in bottom-right corner in Figure 1) about object relationship and orientation, equips the model to answer correctly. The integration enhances the quality of generated responses, especially in tasks that require complex reasoning and context-aware understanding.

In this work, we propose to augment multiple modalities with knowledge graphs to help the model solve complex problems eliciting CoT capabilities. The proposed approach, KAM-CoT, consists of an LM that takes language context, a vision encoder to encode visual features and a graph neural network (GNN) that reasons over the KGs. Following Zhang et al. (2023c), we decouple the reasoning process into two sequential stages. In the first stage, we generate well-reasoned rationales. The second stage takes the generated rationale as an additional input and provides answers. KAM-CoT seamlessly stitches text, vision and graph features together, enabling machines to think and reason coherently, similar to human cognition. We evaluate our proposed model on the ScienceQA (Lu et al. 2022) benchmark. We achieve an average accuracy of 93.87%, surpassing GPT-3.5 (75.17%) by 18% and GPT-4 (83.99%) by 10%. Additionally, KAM-CoT achieves these results with only 280M trainable parameters at a time, demonstrating its cost-efficiency and effectiveness.

This paper makes the following contributions:

1. Graph Extraction: We extract salient triples from ConceptNet (Speer, Chin, and Havasi 2017) based on the given question context.

2. Fusion with KG: We propose a few indicative mechanisms for fusing text and image modalities with the knowledge graph, and examine their efficiency.

3. KAM-CoT: We propose the Knowledge Augmented Multimodal CoT approach, KAM-CoT. The 280M model jointly processes vision, text, and knowledge graph in stages, does step-by-step reasoning to generate plausible reasoning and answers.

We conduct extensive experiments and evaluation on the ScienceQA dataset(Lu et al. 2022), achieving new state-of-the-art performance. We also look into the effects and contributions of each component and discuss potential directions for future research.

Related Work

We explore related works in four key areas: in-context learning, CoT through fine-tuning approaches, vision-language models and knowledge augmented methods.

In-context learning LLMs (Zhao et al. 2023) exhibit the capability of CoT through two principal modes: Zero shot and Few shot. Zero shot performs inference without necessitating any explicit examples or guidance. Recent studies have revealed that LLMs can achieve satisfactory results when prompted with the phrase "Let's think step by step" (Kojima et al. 2022a). In few shot context, LLMs are provided with a set of demonstrative examples that serve as guides, enabling them to grasp and learn patterns from these instances. The examples are curated by human experts.

Auto-CoT introduces the automatic construction of demonstration examples using LLMs (Zhang et al. 2023b). It generates examples with inherent noise. With automatic sampling of diverse questions and post-processing quality control mechanisms, it gets usable chains. Wang et al. (2022a) proposes a decoding self-consistent strategy that samples from a diverse set of reasoning paths and subsequently selects the most consistent answer by marginalizing all possible paths. PROMPTPG (Lu et al. 2023b) employs policy gradient techniques to acquire the ability to discern contextually related examples from the limited set of training samples and then construct the corresponding prompt for a given sample. Chen et al. (2022) proposes *Program of Thoughts*, where the computation is delegated to an interpreter, decoupling complex computation from reasoning and understanding. Another interesting work, *least-to-most prompting* (Zhou et al. 2023) proposes to break a complex problem into simpler ones and solve them sequentially by leveraging the answer from previously solved sub-problems. However, all these approaches are limited to LLMs, reasonably greater than 100B parameters (Wei et al. 2022a).

CoT through fine-tuning approaches Lu et al. (2022) proposes a Science Question-Answer (ScienceQA) dataset that consists of multimodal multiple choice questions with corresponding lectures, explanations and correct answers. Authors observe improvements in question answering by using CoT by 1.20% in few shot GPT-3 and 3.99% in fine-tuned UnifiedQA (Khashabi et al. 2020). MM-CoT (Zhang et al. 2023c) proposes to fine-tune an LM on ScienceQA dataset with CoT method. They propose rationale generation and answer inference in two stages. The model outperforms GPT-3.5 by 16% on this dataset and surpasses human performance.

Vision-Language Models With the proposal of visual question answering tasks (Antol et al. 2015), there have been plenty of works in aligning vision and language modalities. ViLT (Kim, Son, and Kim 2021) proposes a single transformer architecture for text and image modalities that facilitates seamless cross modal interaction. Patch-TRM (Transformer with cross-modal TRM) parses images into ordered patches in a hierarchical pyramid layout (Lu et al. 2021). The patches are encoded with pre-trained ResNet and passed through a vision transformer. VisualBERT proposes a unified architecture that leverages the expressive power of transformer based BERT model and aligns the features extracted from images (Li et al. 2019, 2020). In particular, both visual and textual inputs are masked, and the model learns to predict the masked inputs, enabling it to capture contextual alignment. BLIP2 (Li et al. 2023) proposes QFormer, pretrained with a two-stage strategy to align image encoders and LLMs. Liu et al. (2023b) proposes the Prismer model, that uses an ensemble of domain experts. KOSMOS (Huang et al. 2023) trains a model from scratch on web-scale multimodal corpora, including arbitrarily interleaved text and images, image-caption pairs, and text data.

Recently with the advent of LLaMA models, there has been significant progress in instruction-following language modelling. LLaVA (Liu et al. 2023a) relies on the text-

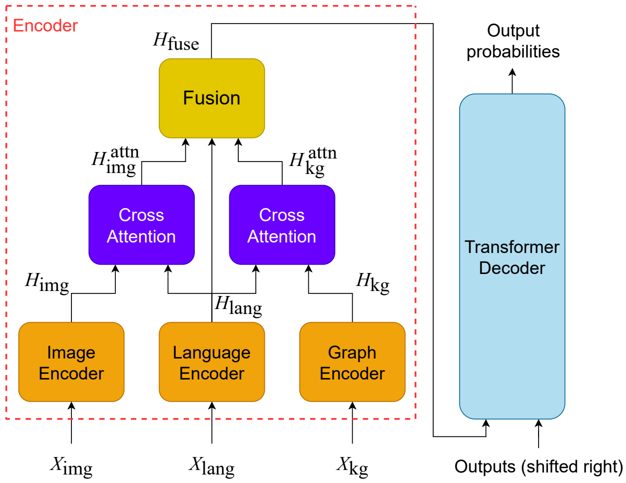

Figure 2: KAM-CoT model architecture.

only GPT-4 (OpenAI 2023) model, to generate multimodal data. The authors propose two stage training: pre-training for feature alignment and instruction-following fine-tuning. LLaMA-Adapter V2 (Gao et al. 2023) proposes a parameter-efficient adapter based visual instruction model that distributes instruction following ability across the entire model. LaVIN (Luo et al. 2023) is another parameter-effecient technique based on mixture of modalities. SCITUNE (Horawalavithana et al. 2023) and T-SciQ (Wang et al. 2023) are science-focused visual and language understanding models. Chameleon (Lu et al. 2023a) mitigates the limitations of accessing up-to-date information, by augmenting LLMs with plug-and-play modules for compositional reasoning. However all these instruction following methods require larger models, usually greater than 7B parameters.

Knowledge augmented methods Several recent studies have explored infusion of structured knowledge into LMs. SKILL (Moiseev et al. 2022) proposes conversion of KG triples into sentences and then using them for pretraining. KagNet (Lin et al. 2019) proposes to ground a question-answer pair from the semantic space to the knowledge-based symbolic space as a schema graph, and then trains a graph convolution network with a hierarchical path-based attention mechanism. QA-GNN (Yasunaga et al. 2021) proposes the use of LMs to estimate the importance of nodes in a KG with respect to the given context, and does joint reasoning over a unified graph. Zhang et al. (2022) proposes the GreaseLM model that fuses encoded representations from pretrained LMs and graph neural networks over multiple layers of language-KG interaction. Extending to multiple modalites, VQA-GNN (Wang et al. 2022b) proposes to unify the image-level scene graph with conceptual knowledge to perform joint reasoning over the unified graph.

Method

We describe the proposed KAM-CoT approach in this section. As an overview, KAM-CoT involves encoding the language, image and the graph input. Note that the graph is derived from the language input. The three modalities are then made to interact with each other using cross-attention. Finally, the fused features are fed to a transformer decoder that generates text autoregressively.

Task Formulation

Given a question q along with k answer choices $\{a_1, a_2, \ldots, a_k\}$, the task is to pick the correct choice. The question q is optionally accompanied by an image X_{img} and a text c that adds context to it.

One potential approach is to use a neural network to generate the right choice directly. However, as already established, chain-of-thoughts reasoning helps in inferring the right answer, especially for complex reasoning tasks (Wei et al. 2022b; Kojima et al. 2022b). We therefore train the model to generate a rationale r for the answer, in the first step. The next step involves picking the correct answer by conditioning the generation process on r, along with the existing inputs. The rationale generation and answer identification models are the same, but they are trained separately from identical initializations. This is similar to the technique used by Zhang et al. (2023c) who deal with just image and text modalities. In our case, we extend their approach to handle graphs as an additional modality that would ground the generation process on factual knowledge.

To obtain the language input for rationale generation, we simply concatenate the different text portions, $X_{\text{lang}}^{\text{rat}} = [q; c; [a_1, a_2, \ldots, a_k]]$. And for answer choice prediction, we append the rationale r as well to obtain $X_{\text{lang}}^{\text{ans}} = [q; c; [a_1, a_2, \ldots, a_k]; r]$.

We extract a subgraph X_{kg} for each sample (discussed in details below). For rationale generation, we learn a model $F_{\text{rat}}(.)$ that generates the rationale r.

$$r = F_{\text{rat}}(X_{\text{lang}}^{\text{rat}}, X_{\text{img}}, X_{\text{kg}}) \tag{1}$$

Similarly, for generating text to identify the right answer, we learn a model $f_{\text{ans}}(.)$.

$$a = F_{\text{ans}}(X_{\text{lang}}^{\text{ans}}, X_{\text{img}}, X_{\text{kg}}) \tag{2}$$

Formalizing the procedure, with the modalities given to the model as input, we compute and maximize the probability of generating the reference text Y, which can either be the rationale or the answer, of length N.

$$p(Y|X_{\text{lang}}, X_{\text{img}}, X_{\text{kg}}) = \prod_{i=1}^{N} p_\theta(Y_i|X_{\text{lang}}, X_{\text{img}}, X_{\text{kg}}, Y_{<i})$$

The model p_θ is made with a combination of a graph encoder and a transformer network. Algorithm 1 lists the steps involved in the KAM-CoT algorithm.

Encode Inputs From Different Modalities

Text Encoding We use a transformer based language encoder to encode X_{lang} to obtain $H_{\text{lang}} = \text{LanguageEncoder}(X_{\text{lang}}) \in \mathbb{R}^{n \times d}$, where n is the number of tokens in X_{lang} and d is the output embedding size of the language encoder.

Algorithm 1: KAM-CoT Reasoning

Input: Language features $X^{\text{rat}}_{\text{lang}}$, Image features X_{img}, and Graph features X_{kg}
Output: Rationale r, Answer a

Construct input $X = \{X^{\text{rat}}_{\text{lang}}, X_{\text{img}}, X_{\text{kg}}\}$
$r \leftarrow F_{\text{rat}}(X)$
Concatenate r to $X^{\text{rat}}_{\text{lang}}$, to make $X^{\text{ans}}_{\text{lang}} \leftarrow [X^{\text{rat}}_{\text{lang}}; r]$
Construct new input $X' = \{X^{\text{ans}}_{\text{lang}}, X_{\text{img}}, X_{\text{kg}}\}$
$a \leftarrow F_{\text{ans}}(X')$
procedure $F(X)$
    Get the encoded representations, H_{lang}, H_{img}, and H_{kg}
    Obtain the feature representations, $H^{\text{attn}}_{\text{img}}$, and $H^{\text{attn}}_{\text{kg}}$
    Fuse these representations with H_{lang} to obtain H_{fuse}
    Input H_{fuse} to the decoder to get the target Y
    return Y
end procedure

Image Encoding We encode the image X_{img} using a transformer based image encoder to obtain $H_{\text{img}} = \text{ImageEncoder}(X_{\text{img}})W_{\text{img}} \in \mathbb{R}^{m\times d}$ where m is the number of patches in the image. The projection matrix W_{img} brings the output embedding dimension to d, same as that of H_{lang}.

Subgraph Extraction For every sample, we extract a subgraph from ConceptNet (Speer, Chin, and Havasi 2017) by following a method similar to that in Yasunaga et al. (2021). We group the relations in ConceptNet into 17 distinct types. These relations can be either forward or backward, yielding a total of 34 possible edge types. The triples are converted to sentences, and corresponding sentence patterns are stored. These patterns are used to ground and extract nodes from the question, context and answer choices. A subgraph is made of (i) $\mathbf{V}$, a set of nodes, (ii) $\mathbf{E}$, a set of edges and (iii) $\boldsymbol{\phi}$, a function which maps every edge to an integer in the range $[0, 33]$, representing the edge type. To get the initial node embeddings, we the same pretrained checkpoint of the language encoder used for text encoding, and average the embeddings over the span of all occurences of that node (Feng et al. 2020). The thought behind using the same language encoder checkpoint is to ensure that the language and node embeddings start from the same space.[1] Let N_{qa} represent this set of grounded nodes. For every pair of nodes, $n_{\text{a}}, n_{\text{b}} \in N_{\text{qa}}$, we append all common nodes in their 1-hop neighbourhood into $N_{\text{1-hop}}$. We repeat this process for each pair of nodes in N_{qa} and $N_{\text{1-hop}}$ and append the nodes into $N_{\text{2-hop}}$. This way, we get a graph connecting all nodes in N_{qa} to each other with a path length of atmost 2 intermediate nodes: $\mathbf{V} = N_{\text{qa}} \cup N_{\text{1-hop}} \cup N_{\text{2-hop}}$. Since the number of nodes could grow exponentially, we follow the pruning strategy in Yasunaga et al. (2021) to keep the top 200 nodes for every sample. For the edges, we build an embedding table and learn embeddings during training.

Graph Encoding Using a combination of graph layers, we encode the extracted subgraph X_{kg} to obtain the node embeddings $H_{\text{kg}} = \text{KGEncoder}(X_{\text{kg}}) \in \mathbb{R}^{p\times d}$, where p is the number of extracted nodes.

Interaction Between Modalities

We use cross-attention to enable the interaction between the representations of text, image and subgraph. For this we use two seperate single-headed attention modules (see Figure 2). For the first attention module, the language and image embeddings interact. Similarly, in another attention module interaction between language and node embeddings happen.

$$H^{\text{attn}}_{\text{img}} = \text{softmax}\Big(\frac{H_{\text{lang}}H^{\top}_{\text{img}}}{\sqrt{d}}\Big)H_{\text{img}} \tag{3}$$

$$H^{\text{attn}}_{\text{kg}} = \text{softmax}\Big(\frac{H_{\text{lang}}H^{\top}_{\text{kg}}}{\sqrt{d}}\Big)H_{\text{kg}} \tag{4}$$

Fusion

We use gated fusion (Wu et al. 2021; Zhang and Zong 2020; Li et al. 2022) to get the final representation.

$$\begin{aligned} S_\alpha &= H_{\text{lang}}W_1 + H^{\text{attn}}_{\text{img}}W_2 + H^{\text{attn}}_{\text{kg}}W_3 \in \mathbb{R}^{n\times d} \\ S_\beta &= H_{\text{lang}}W_4 + H^{\text{attn}}_{\text{img}}W_5 + H^{\text{attn}}_{\text{kg}}W_6 \in \mathbb{R}^{n\times d} \\ S_\gamma &= H_{\text{lang}}W_7 + H^{\text{attn}}_{\text{img}}W_8 + H^{\text{attn}}_{\text{kg}}W_9 \in \mathbb{R}^{n\times d} \end{aligned} \tag{5}$$

$$\begin{aligned} \alpha_{ij}, \beta_{ij}, \gamma_{ij} &= \text{softmax}([S_{\alpha_{ij}}, S_{\beta_{ij}}, S_{\gamma_{ij}}]) \\ H_{\text{fuse}} &= \alpha \cdot H_{\text{lang}} + \beta \cdot H^{\text{attn}}_{\text{img}} + \gamma \cdot H^{\text{attn}}_{\text{kg}} \in \mathbb{R}^{n\times d} \end{aligned} \tag{6}$$

Here $\alpha, \beta, \gamma \in [0, 1]^{n\times d}$ and sum to 1 element-wise, and all $W \in \mathbb{R}^{d\times d}$. We will refer to this fusion method as **Fusion-1**. We discuss and compare a few other fusion variants in the Discussion and Analysis section.

Decoding

We use a transformer decoder that utilizes H_{fuse} to generate text autoregressively.

$$p(Y_t|Y_{<t}, X_{\text{lang}}, X_{\text{img}}, X_{\text{kg}}) = \text{Decoder}(Y_{<t}, H_{\text{fuse}}) \tag{7}$$

Experiments

Dataset

We evaluate our method on the ScienceQA benchmark (Lu et al. 2022). It comprises of 21208 multiple-choice questions with multimodal contexts, sourced from the science curriculum. It covers substantial domain diversity, spanning 3 subjects, 26 topics, 127 categories and 379 skills. ScienceQA provides us with an in-house training, dev and test split containing 12726, 4241 and 4241 samples respectively.

Baseline Comparisons

We choose the following baselines, (i) VQA models (Kim, Son, and Kim 2021; Lu et al. 2021; Li et al. 2020), (ii) Models with similar backbones (Khashabi et al. 2020; Lu et al. 2022; Zhang et al. 2023c; Wang et al. 2023), (iii) Parameter-efficient finetuned LLMs (Zhang et al. 2023a; Luo et al. 2023), and (iv) the GPT family and GPT-assisted models (OpenAI 2022, 2023; Liu et al. 2023a).

[1]We also experiment with using image captions for grounding. In that case, we simply append the caption to the existing context.

Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Human Average	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
ViLT (Kim, Son, and Kim 2021)	112M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM (Lu et al. 2021)	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT (Li et al. 2020)	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA$_{\text{Base}}$ (Khashabi et al. 2020)	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
UnifiedQA$_{\text{Base}}$ w/ CoT (Lu et al. 2022)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
MM-CoT$_{\text{T5-Base}}$ (Zhang et al. 2023c)	223M	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{\text{FLAN-T5-Base}}$	250M	91.5	74.92	90.09	91.69	84.28	90.52	88.14	87.01	87.74
MM-T-SciQ$_{\text{Base}}$ (Wang et al. 2023)	223M	91.52	91.45	92.45	91.94	90.33	92.26	92.11	91.10	91.75
LLaMA-Adapter (Zhang et al. 2023a)	6B (1.2M)	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LaVIN-13B (Luo et al. 2023)	13B (5.4M)	89.88	94.49	89.92	88.95	87.61	91.85	91.45	89.72	90.83
GPT-3.5 w/ CoT (OpenAI 2022)	>175B	75.44	70.97	78.09	74.68	67.43	79.93	78.23	69.68	75.15
GPT-4 w/ CoT (OpenAI 2023)	>175B	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04	83.99
LLaVa (GPT-4) (Liu et al. 2023a)	13B	91.56	**96.74**	91.09	90.62	88.99	93.52	92.73	92.16	92.53
KAM-CoT$_{\text{T5-Base}}$ **(Ours)**	223M	93.21	92.21	90.64	93.21	**93.26**	91.50	92.51	92.42	92.48
KAM-CoT$_{\text{FLAN-T5-Base}}$ **(Ours)**	250M	**94.76**	92.24	**93.36**	**94.53**	93.16	**94.15**	**94.24**	**93.21**	**93.87**

Table 1: Comparing the results against baselines. Here, Size = size of the backbone model, NAT = Natural Science, SOC = Social Science, LAN = Language Science, TXT = Text context, IMG = Image context, NO = No context, G1-6 = Grade 1 to 6, G7-12 = from Grade 7 to 12. Segment 1 compares against the human average. Segment 2 shows the performance of chosen VQA baselines. Segment 3 has models whose backbone sizes are comparable to ours. In Segment 4, we show parameter-efficient finetuned versions of larger models, and the number of trainable parameters are provided inside parantheses. Segment 5 has the performance of the GPT family. MM-CoT$_{\text{FLAN-T5-Base}}$ here has been given caption as context along with the vision features. Results, other than ours and MM-CoT$_{\text{FLAN-T5-Base}}$, are taken from respective papers and the ScienceQA leaderboard.

Training Details

The size of the proposed model is 254M with T5-Base and 280M with FLAN-T5-Base. All our experiments are run on a single NVIDIA A100 40G GPU. We train our models for 20 epochs, and also evaluate them after each, with ScienceQA's dev split. We use a learning rate of 5e-5 and batch-size of 1, a maximum input length of 512 tokens, and maximum output length of 512 and 64 tokens for rationale and answer generation respectively.

Experimental Setup

For our experiments, we discuss the effect of using different image encoders. (i) CLIP (Radford et al. 2021) aligns images and text into a common embedding space. (ii) DETR (Carion et al. 2020) leverages transformers to perform object detection and localization. The chosen variants of DETR[2] and CLIP[3] are used without their classification heads, to provide patch embeddings of shape $(100, 256)$ and $(49, 2048)$, respectively.

We experiment with caption features as well, where captions are generated using ViT-GPT2.[4] Yet another set of experiments use these captions for extracting graph nodes. In this case, right after generating the possible entailments of the sample, we put the caption seperated by a white-space. The grounding process then continues as discussed in the Method section. We also experiment with both the above mentioned settings.

To encode the knowledge-graph we use two layers: a Relational Graph Attention layer (Busbridge et al. 2019), followed by a Graph Convolutional layer (Kipf and Welling 2017), both implemented in PyTorch Geometric (Fey and Lenssen 2019). We refrain from using more than two graph layers as that might lead to a node forgetting its own identity (Li, Han, and Wu 2018). The first graph layer uses 768 input and output features, matching the language encoder's embedding dimension size. It is also provided with the number of possible relations, 34 and the edge embedding size, 64. Next, the Graph Convolution layer is given only the input and output feature sizes, both being set at 768. As mentioned in the Method section, for representing the edges, we learn an embedding table in the training process. Given an integer for the edge-type, it produces an embedding, $e_{edge} \in \mathbb{R}^{64}$ for that edge, and is fed to the graph-encoder.

Our approach uses T5-Base (Raffel et al. 2020) as its backbone. The well defined encoder-decoder architecture gives a good entry-point to introduce other modalities. To ensure the applicability of our approach to other language models, we conduct experiments and present results on the instruction-tuned FLAN-T5-Base (Chung et al. 2022) also.

Results

To assess the effectiveness of our model, we use two evaluation metrics: average accuracy and RougeL (Lin 2004). Average accuracy quantifies the model's correctness in predicting the correct answer, and is treated as the primary metric for evaluating the quality of our method. We use the RougeL metric to to compare the generated rationale to the human reference, as done in Zhang et al. (2023c). ScienceQA contains multiple groups, that enables us to compare group-wise accuracies, giving an insight to the model's strengths and limitations within each group, which is valuable in understanding how the model generalizes across content areas.

[2]https://huggingface.co/facebook/detr-resnet-101-dc5
[3]https://huggingface.co/google/vit-base-patch16-384
[4]https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

Image Features	Feature Size	RougeL	Avg. Acc
DETR	(100, 256)	**98.29**	**91.65**
CLIP	(49, 2048)	98.15	91.02

Table 2: Comparative results using different image encoders with T5-Base. DETR outperforms the CLIP based encoding.

Method	RougeL	Avg. Acc
without captions	98.29	91.65
with captions as context	98.33	92.45
captions for node extraction	98.31	91.84
captions for nodes + context	**98.32**	**92.48**

Table 3: Summary of results that showcase different approaches using captions with T5-Base.

For a fair and consistent evaluation, we obtain the scores of the baseline models directly from their respective research papers. Additionally, we take scores from the ScienceQA leaderboard[5] for closed-source models. This enables us to make informed assessments of our model's contributions in comparison to existing state-of-the-art. Table 1 shows the main results. Our model outperforms all other known approaches under 300M and does not use any very large auxiliary model. With FLAN-T5-Base as the backbone, we achieve a RougeL score of 98.40 and an average accuracy of 93.87, which is well above the performance of GPT3.5 (75.17%), and also surpasses LLaVa (92.53%) by 1.34%. This conceretely establishes that our proposed method is superior compared to other approaches including LLMs, while being under 300M parameters.

A closer look into Table 1 reveals that questions about Natural Science, Social Science and Language Science see a boost compared to the baselines. The same is also observed for No-Context questions. ConceptNet is expected to aid with these kind of questions, which is visible here clearly.

We conduct further experiments and ablation studies to delve deeper into the performance and robustness of our proposed model. We also explore the effects of varying the individual modalities and encoders. We explore more fusion methods in the Additional Fusion Mechanisms subsection.

Unless explicitly mentioned, all experiments are trained and evaluated for 20 epochs, and then tested on the test-split.

Table 2 shows the effect of using different image encoders. DETR gives a marginal improvement (0.63%) over CLIP features, despite having a smaller feature size (74k floats lesser) per sample, making it our default choice.

We observe from Table 3, captions concatenated with the context gave a boost to both the rationale and the accuracy scores. In another setting where captions are concatenated with the context, and then used to extract nodes, shows a marginal boost over not using them at all $(91.65 \rightarrow 91.84)$, but also with a very little fall in the RougeL score (0.02).

The final combination, where captions are added to the context and also used for extracting node embeddings, turns out to be the best setting for average accuracy.

[5]https://scienceqa.github.io/leaderboard.html

Number of nodes	RougeL	Avg. Acc
50	97.78	88.66
100	97.84	88.85
200	**97.85**	**89.51**

Table 4: Effect of varying the number of nodes in a graph with T5-Base as the backbone.

Image	Captions	KG	RougeL	Avg. Acc
✗	✗	✗	-	83.42[6]
✓	✗	✗	-	85.85 [6]
✓	context	✗	97.27	87.74
✓	context	✓	98.34	92.62
✓	**context, nodes**	✓	**98.40**	**93.87**

Table 5: Ablation study on the KAM-CoT framework, using FLAN-T5-Base.

We study the effect of taking the top 50, 100 and 200 nodes. If the node extraction process yields a smaller number of nodes, they are zero-padded to the minimum number. To expedite these experiments with varying number of nodes, and to reduce GPU consumption, we limit training to 10 epochs. Limiting the maximum number of nodes has a proportional effect on the accuracy. Table 4 shows the trend that more nodes help the model reason and choose better. Although we could not perform exhaustive experiments with higher number of nodes, we anticipate that the performance would saturate and might even decline beyond a certain threshold. We defer this aspect to future research.

Having explored the effects of various settings over the modalities, we perform ablation studies, with FLAN-T5-Base as the backbone. The complete model amounts to a total of 279M trainable parameters with the graph encoder included. From Table 5, it is easily seen that just plugging in the graph encoder gives an accuracy boost of 4.88%, totaling to 92.62, which surpasses the performance of LLaVA (Table 1) with 13B parameters, and is yet not the highest score we could reach.

As reported in the beginning of this section, the best out of all our experiments come with the captions as context + node extraction setting. With 280M paramters, our achitecture has a RougeL score of 98.40 and an average accuracy of 93.87, with a model **47 times smaller** than its next best performer.

Discussion and Analysis

In this section, we examine, a few alternative fusion mechanisms, model convergence, and results using subset of train data.

Additional Fusion Mechanisms

Unlike the bottleneck-style (Yasunaga et al. 2021) interaction between node embedding and other modalities, our fusion mechanisms have no such constraints. Along with the proposed primary fusion method in the Fusion subsection, we experiment with two more settings.

[6]Results are taken from (Zhang et al. 2023c)

Fusion Method	# Parameters	RougeL	Avg. Acc
Fusion-1	254M	**98.29**	**91.65**
Fusion-2	251M	98.23	91.23
Fusion-3	250M	98.14	90.14

Table 6: Comparative performance of the varying fusion methods. Fusion-1 outperforms the other fusion methods.

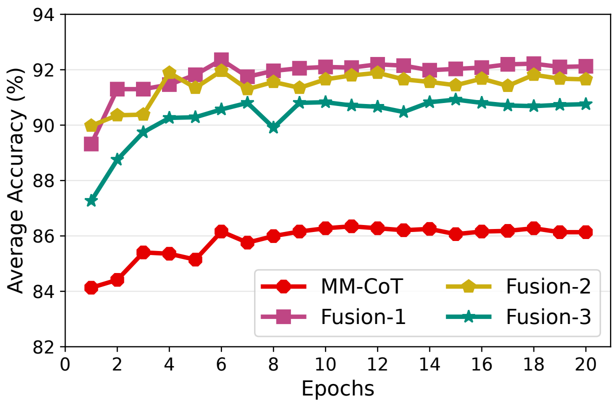

Figure 3: Performance of the fusion mechanisms on the validation set, evaluated using T5-Base.

2-step fusion (Fusion-2) In the first stage, we fuse language-vision and language-KG features and get $H_{\text{img,kg}}$. Considering language as the primarily modality, we fuse it with $H_{\text{img,kg}}$ in the second stage.

$$\begin{aligned} \lambda_a &= \text{sigmoid}(H_{\text{img}}^{\text{attn}} W_1 + H_{\text{kg}}^{\text{attn}} W_2) \in \mathbb{R}^{n \times d} \\ H_{\text{img,kg}} &= (1 - \lambda_a) \cdot H_{\text{img}}^{\text{attn}} + \lambda_a \cdot H_{\text{kg}}^{\text{attn}} \in \mathbb{R}^{n \times d} \end{aligned} \tag{8}$$

$$\begin{aligned} \lambda_b &= \text{sigmoid}(H_{\text{img,kg}} W_3 + H_{\text{lang}} W_4) \in \mathbb{R}^{n \times d} \\ H_{\text{fuse}} &= (1 - \lambda_b) \cdot H_{\text{img,kg}} + \lambda_b \cdot H_{\text{lang}} \in \mathbb{R}^{n \times d} \end{aligned} \tag{9}$$

1-step fusion (Fusion-3) In this approach we take the linear projection of H_{lang}, $H_{\text{img}}^{\text{attn}}$, $H_{\text{kg}}^{\text{attn}}$ and compute their weighted sum to merge all the modalities.

$$S_\alpha = H_{\text{lang}} W_1 \ , \ S_\beta = H_{\text{img}}^{\text{attn}} W_2 \ , \ S_\gamma = H_{\text{kg}}^{\text{attn}} W_3 \ , \tag{10}$$

$$\begin{aligned} \alpha_{ij}, \beta_{ij}, \gamma_{ij} &= \text{softmax}([S_{\alpha_{ij}}, S_{\beta_{ij}}, S_{\gamma_{ij}}]) \\ H_{\text{fuse}} &= \alpha \cdot H_{\text{lang}} + \beta \cdot H_{\text{img}}^{\text{attn}} + \gamma \cdot H_{\text{kg}}^{\text{attn}} \in \mathbb{R}^{n \times d} \end{aligned} \tag{11}$$

We summarise the results of these fusion mechanisms in Table 6 and find that Fusion-1 gives the best performance on ScienceQA test data.

Comparing Model Convergence

Figure 3 compares our model's convergence trend (with all fusion techniques) with MM-CoT (Zhang et al. 2023c) on the validation. We observe that the proposed method as well as MM-CoT converge at 10 epochs. Note that, the accuracy of the proposed approach starts much higher as compared to MM-CoT. Also, Fusion-1 demonstrates the highest accuracy, along with greater stability in comparison to others.

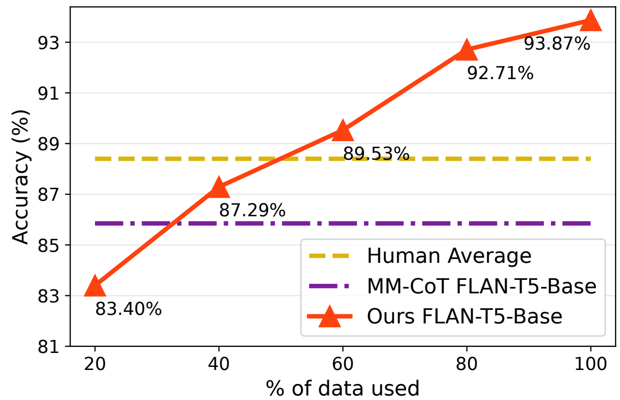

Figure 4: Comparative performance using subsets of training data with MM-CoT$_{\text{FLAN-T5-Base}}$ (100% training data, Zhang et al. (2023c)), and the human average.

Dataset Variation

To examine the scalability of the proposed model, we also train on subsets of the training data. These sets are made in the proportion of 20%, 40%, 60% and 80% of all the 12k total training samples, preserving the distribution over the 26 topics. Figure 4 shows that KAM-CoT surpasses human accuracy (88.4%) even when trained with only 50% of the training data. Surprisingly, the model outperforms the fully trained MM-CoT (Flan-T5$_{\text{Base}}$) (93.87% vs 85.85%) with only 35% of the training data. The results highlight the model's generalization ability with little training data.

We also evaluate the model with A-OKVQA dataset. The proposed model outperforms the baseline by 3.67%.

Conclusion

In this paper, we propose KAM-CoT, Knowledge Augmented Multimodal Chain of Thought reasoning, to enhance the reasoning capability and quality of answers from language models. We propose a framework that uses CoT reasoning, leverages knowledge graphs and other modalities for a comprehensive understanding of multimodal tasks. We provide a few possible methods to fuse these modalities. We find that the incorporation of KG in the two-stage training process helps reduce hallucinations. With only 280M parameters at a time, our approach yields a new state-of-the-art having an accuracy 93.87%, outperforming GPT-3.5 by 18% by and GPT-4 by 10%. In the future, we want to further integrate specific knowledge-intensive domains, and also explore efficient fusion mechanisms. We would also like to scale our solution to larger models like the LLaMA family.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Busbridge, D.; Sherburn, D.; Cavallo, P.; and Hammerla, N. Y. 2019. Relational Graph Attention Networks. arXiv:1904.05811.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, 213–229. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58451-1.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

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. arXiv:2210.11416.

Feng, Y.; Chen, X.; Lin, B. Y.; Wang, P.; Yan, J.; and Ren, X. 2020. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1295–1309. Online: Association for Computational Linguistics.

Fey, M.; and Lenssen, J. E. 2019. Fast Graph Representation Learning with PyTorch Geometric. arXiv:1903.02428.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Horawalavithana, S.; Munikoti, S.; Stewart, I.; and Kvinge, H. 2023. SCITUNE: Aligning Large Language Models with Scientific Multimodal Instructions. *arXiv preprint arXiv:2307.01139*.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Liu, Q.; Aggarwal, K.; Chi, Z.; Bjorck, J.; Chaudhary, V.; Som, S.; Song, X.; and Wei, F. 2023. Language Is Not All You Need: Aligning Perception with Language Models. *ArXiv*, abs/2302.14045.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1896–1907. Online: Association for Computational Linguistics.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *International Conference on Machine Learning*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022a. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022b. Large Language Models are Zero-Shot Reasoners. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 22199–22213. Curran Associates, Inc.

Li, B.; Lv, C.; Zhou, Z.; Zhou, T.; Xiao, T.; Ma, A.; and Zhu, J. 2022. On Vision Features in Multimodal Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6327–6337. Dublin, Ireland: Association for Computational Linguistics.

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. *arXiv preprint arXiv:2301.12597*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2020. What Does BERT with Vision Look At? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5265–5275. Online: Association for Computational Linguistics.

Li, Q.; Han, Z.; and Wu, X.-m. 2018. Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2829–2839. Hong Kong, China: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. arXiv:2304.08485.

Liu, S.; Fan, L.; Johns, E.; Yu, Z.; Xiao, C.; and Anandkumar, A. 2023b. Prismer: A Vision-Language Model with An Ensemble of Experts. In *ArXiv Preprint*.

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. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Lu, P.; Peng, B.; Cheng, H.; Galley, M.; Chang, K.-W.; Wu, Y. N.; Zhu, S.-C.; and Gao, J. 2023a. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.

Lu, P.; Qiu, L.; Chang, K.-W.; Wu, Y. N.; Zhu, S.-C.; Rajpurohit, T.; Clark, P.; and Kalyan, A. 2023b. Dynamic Prompt Learning via Policy Gradient for Semi-structured Mathematical Reasoning. In *The Eleventh International Conference on Learning Representations*.

Lu, P.; Qiu, L.; Chen, J.; Xia, T.; Zhao, Y.; Zhang, W.; Yu, Z.; Liang, X.; and Zhu, S.-C. 2021. IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning. *ArXiv*, abs/2110.13214.

Luo, G.; Zhou, Y.; Ren, T.; Chen, S.; Sun, X.; and Ji, R. 2023. Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models. arXiv:2305.15023.

Moiseev, F.; Dong, Z.; Alfonseca, E.; and Jaggi, M. 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*, 1581–1588. Seattle, United States: Association for Computational Linguistics.

OpenAI. 2022. ChatGPT, OpenAI. chat.openai.com. Accessed: 2023.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

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. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, L.; Hu, Y.; He, J.; Xu, X.; Liu, N.; Liu, H.; and Shen, H. T. 2023. T-SciQ: Teaching Multimodal Chain-of-Thought Reasoning via Large Language Model Signals for Science Question Answering. arXiv:2305.03453.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022a. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wang, Y.; Yasunaga, M.; Ren, H.; Wada, S.; and Leskovec, J. 2022b. Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. *arXiv preprint arXiv:2205.11501*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022a. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. Survey Certification.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, Z.; Kong, L.; Bi, W.; Li, X.; and Kao, B. 2021. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6153–6166. Online: Association for Computational Linguistics.

Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 535–546. Online: Association for Computational Linguistics.

Zhang, J.; and Zong, C. 2020. Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63: 2028 – 2050.

Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; and Qiao, Y. 2023a. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. arXiv:2303.16199.

Zhang, X.; Bosselut, A.; Yasunaga, M.; Ren, H.; Liang, P.; Manning, C. D.; and Leskovec, J. 2022. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. arXiv:2201.08860.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2023b. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023c. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv:2302.00923.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*.