Zero-shot Commonsense Reasoning over Machine Imagination

Anonymous ACL submission

Abstract

Recent approaches to zero-shot commonsense reasoning have enabled Pre-trained Language Models (PLMs) to learn a broad range of commonsense knowledge without being tailored to specific situations. However, they often suffer from human reporting bias inherent in textual commonsense knowledge, leading to discrepancies in understanding between PLMs and humans. In this work, we aim to bridge this gap by introducing an additional information channel to PLMs. We propose IMAGINE (Machine **Imagin**ation-based **Re**asoning), a novel zero-shot commonsense reasoning framework designed to complement textual inputs with visual signals derived from machine-generated images. To achieve this, we enhance PLMs with imagination capabilities by incorporating an image generator into the reasoning process. To guide PLMs in effectively leveraging machine imagination, we create a synthetic pre-training dataset that simulates visual question-answering. Our extensive experiments on diverse reasoning benchmarks and analysis show that IMAGINE outperforms existing methods by a large margin, highlighting the strength of machine imagination in mitigating reporting bias and enhancing generalization capabilities[1].

1 Introduction

Commonsense reasoning has been considered a crucial milestone in the pursuit of artificial general intelligence (Gunning, 2018). While Pre-trained Language Models (PLMs; Devlin et al., 2019; Brown et al., 2020) often exhibit near-human reasoning capabilities after being fine-tuned on specific commonsense datasets, they face challenges in zero-shot scenarios where examples differ significantly from their training data distribution (Mitra et al., 2019; Kim et al., 2022). Overcoming this limitation

[1]Our code and data are available at https://anonymous.4open.science/r/Imagine-C35A

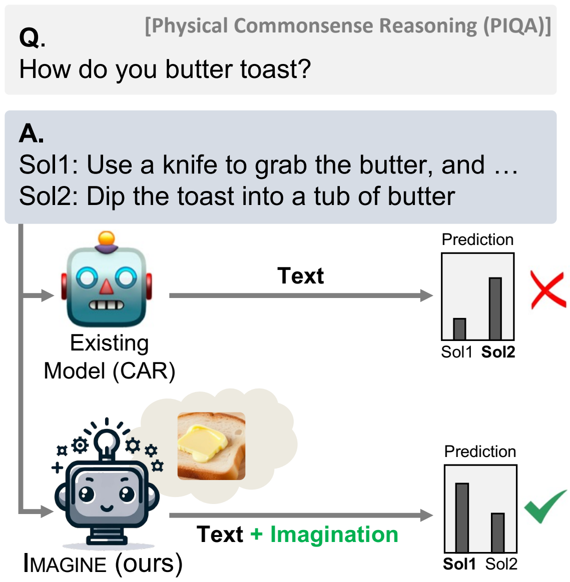


Figure 1: Example from the PIQA (Bisk et al., 2020) with model predictions. Compared to the existing methods, IMAGINE performs reasoning with imagination.

is crucial for achieving human-level proficiency in natural language understanding.

One promising approach to this limitation is injecting commonsense knowledge from external Knowledge Bases (KBs; Sap et al., 2019a; He et al., 2022b) into PLMs. Specifically, this involves transforming knowledge entities into a question-answering (QA) format, resulting in a synthetic QA dataset. This constructed dataset is then used to train PLMs similarly to the pre-training phase. Since the knowledge bases can cover a wide spectrum of commonsense knowledge, this approach leads to substantial improvements in reasoning ability across diverse situations without specializing in specific knowledge (Wang et al., 2023, 2024).

However, they often suffer from human reporting bias (Gordon and Durme, 2013), as textual commonsense knowledge only captures the most frequently occurring scenarios, thereby neglecting less common but equally critical knowledge necessary for comprehensive reasoning. Figure 1 illustrates a case where a recent model (Wang et al., 2023) fails to accurately reason about the question "*How do you butter toast?*". Since the existing models rely solely on textual inputs, they often neglect contextual details, such as the fact that butter is typically too solid to be dipped. In contrast, humans can easily answer such questions by visually imagining the shape, solidity, and interactions of butter with other objects. This observation motivates us to explore additional modalities to complement textual commonsense knowledge.

In this paper, we introduce IMAGINE (Machine **Imagin**ation-based **Re**asoning), a novel zero-shot commonsense reasoning framework designed to circumvent the reporting bias inherent in textual inputs. Inspired by the cognitive studies highlighting the beneficial effects of visual imagery on language understanding (Gambrell and Bales, 1986; Dessalegn and Landau, 2013), IMAGINE is designed to leverage visual signals to complement textual inputs. To achieve this, we integrate PLMs with a conditional image generator, enabling machine imagination capabilities. To guide the model in learning to utilize visual and textual inputs jointly, we create a synthetic VQA dataset, which is then used to optimize PLMs. By acquiring a broad spectrum of commonsense knowledge along with visual signals, IMAGINE enhances reasoning capabilities while circumventing human reporting bias.

To verify the effectiveness of IMAGINE, we perform extensive experiments, encompassing diverse reasoning benchmarks, architectures, and scales. The experimental results convincingly demonstrate that IMAGINE surpasses existing methods, including large language models, in reasoning capabilities. Moreover, our in-depth analysis reveals that IMAGINE effectively enables PLMs to adaptively leverage machine imagination capabilities in a beneficial manner. The contributions of this paper include the following:

- We introduce IMAGINE, a novel zero-shot commonsense reasoning framework, aimed at mitigating reporting bias and enhancing the generalizability of PLMs.
- We construct a synthetic VQA dataset to enable PLMs to jointly utilize textual and visual signals while achieving commonsense reasoning ability.
- We demonstrate that IMAGINE surpasses state-of-the-art zero-shot reasoning models across diverse reasoning tasks, highlighting the significance of machine imagination.

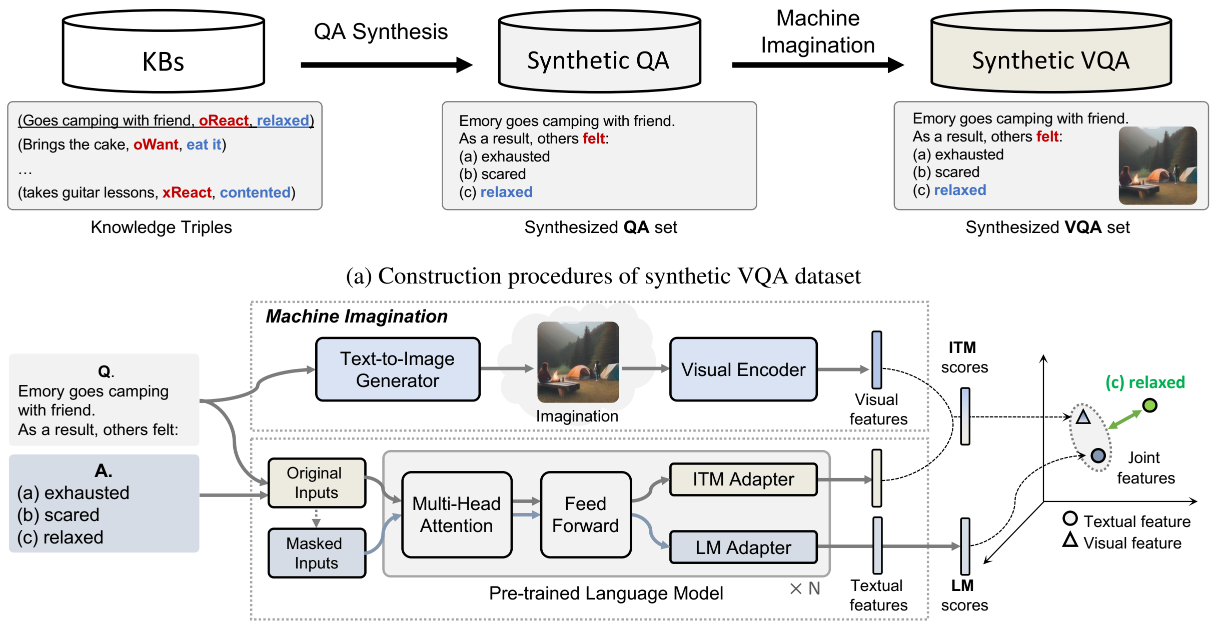
2 Related Work

2.1 Zero-shot Commonsense Reasoning

There are two major approaches to zero-shot commonsense reasoning. The first approach involves utilizing the inherent capabilities of the off-the-shelf PLMs without updating their parameters. For example, Trinh and Le (2018) utilized the perplexity of vanilla language modeling, and Li et al. (2022) leveraged PLMs with specifically-designed prompting. Shwartz et al. (2020) solicited the commonsense knowledge from the language models through an iterative self-talk. Similarly, Dou and Peng (2022) obtained additional knowledge for reasoning based on the cloze-style translation. The second approach involves leveraging external commonsense knowledge bases (e.g., ATOMIC (Sap et al., 2019a), ConceptNet (Speer et al., 2017)) to provide language models with additional knowledge. Specifically, recent studies have transformed the knowledge entities (e.g., triplets of (head, relation, tail)) into synthetic QA pairs and trained the models with them (Banerjee and Baral, 2020; Ma et al., 2021). Recently, Wang et al. (2023) further improved the synthetic signals through a conceptualization process (Song et al., 2011) which abstracts a commonsense knowledge triplet to many higher-level instances. Subsequently, Wang et al. (2024) injected the instantiation phase into the process of synthetic dataset generation with the help of the generation capabilities of LLMs.

2.2 Visual Information for Natural Language Understanding

A few previous works have leveraged machine imagination to address Natural Language Understanding (NLU) problems. For example, Tan and Bansal (2020) proposed VOKEN, which introduces visual supervision into language model pre-training by incorporating external knowledge from images retrieved for the tokens. Instead of retrieving visual information, Lu et al. (2022) proposed generating synthetic images (i.e., imagination) based on a generative model to tackle downstream NLU tasks. In the context of commonsense reasoning, Liu et al. (2022) utilized visual information to comprehend spatial commonsense knowledge (e.g., *how big is a*



(a) Construction procedures of synthetic VQA dataset

(b) Inference and optimization procedures of IMAGINE (ours)

Figure 2: Overall procedures for (a) constructing a synthetic VQA dataset and (b) the inference/optimization phase of IMAGINE (ours) using the given QA pair. The process starts with the textual pair consisting of a question and its answers, followed by the generation of visual signals (i.e., imagination) conditioned on the question. The two distinct features from visual and textual models are then utilized to derive a comprehensive prediction.

lion?). Similar to the proposed method, Yang et al. (2022) introduced Z-LaVI, which integrated visual information with PLMs through both retrieval and synthesis to achieve zero-shot reasoning abilities.

3 Machine Imagination-based Reasoning

In this section, we elaborate on the proposed method, namely IMAGINE (Machine **Imagin**ation-based R**e**asoning), for zero-shot commonsense reasoning. The core strategy is to complement textual commonsense knowledge with visual signals derived from machine-generated images. To achieve this, we first couple the PLMs with a text-to-image generator (§3.1), enabling machine imagination in text-based PLMs. We then construct a large-scale synthetic VQA dataset to learn the joint use of textual and visual signals in the reasoning process (§3.2). By optimizing the model with additional signals that encapsulate commonsense knowledge, IMAGINE can effectively perform commonsense reasoning while avoiding human reporting bias inherent in textual inputs (§3.3, §3.4). The overall procedure is depicted in Figure 2.

3.1 Machine Imagination in PLMs

We start by introducing the machine imagination in text-based PLMs. We denote PLMs as $\mathcal{M}_T$, which serve as the backbone for zero-shot commonsense reasoning. For machine imagination, we incorporate two additional models to process visual signals. Specifically, we introduce: (i) a text-to-image generator, $\mathcal{M}_{T2I}$, which creates relevant images by conditioning the textual inputs, and (ii) a visual encoder, $\mathcal{M}_I$, which acts as a feature extractor for the given images.

The overall mechanism of machine imagination operates as follows: Given a textual input, the text-to-image model $\mathcal{M}_{T2I}$ initially generates an image that captures the essence of the text. With these generated images linked to textual inputs, both PLMs, $\mathcal{M}_T$, and the visual encoder, $\mathcal{M}_I$, jointly encode the textual input and the generated image. The resultant features are then utilized to derive the comprehensive predictions.

3.2 Synthetic VQA Construction

Following the previous works (Ma et al., 2021; Wang et al., 2023), we achieve zero-shot commonsense reasoning ability by constructing the synthetic QA dataset from the knowledge base. On top of this dataset, we build a synthetic visual question-answering (synthetic VQA) dataset with the help of machine imagination. The dataset is designed to: (i) instill commonsense reasoning abilities in

	Q: Emory is walking home. Emory is seen as... A1: Bossy **A2: Tired** A3: Independent		Q: A group of people walking down a street. Where is this scene from? A1: This scene takes place in a university **A2: It looks like the middle east** A3: This scene is set before the nineteen hundreds
	Q: Berkeley folds his tent. Berkeley is seen as... A1: Withdrawn A2: Dedicated **A3: Adventurous**		Q: A man and a woman sitting at a bar. Is Sam currently drunk? A1: Yes, Bali recently drank alcohol **A2: Yes, Sam is intoxicated** A3: Possibly, but not presently

Figure 3: Examples of the Synthetic VQA dataset. The examples on the left are sourced from AbstractATOMIC (Wang et al., 2023), while the two examples on the right are sourced from VCR (Zellers et al., 2019). **Bold** indicates the correct answer, and underline denotes the generated image caption.

PLMs and (ii) teach them to harmoniously utilize both textual and visual inputs.

The objective of this process is to construct VQA pairs (Q, A, I), where each pair includes a natural language question Q, a set of n answer choices $A = A_1, A_2, ..., A_n$, including one ground-truth answer and $n-1$ distractors, along with an image I that corresponds to the question.

Synthetic QA We first construct textual QA pairs from the KBs by following the recent work (Wang et al., 2023). Specifically, we transform the knowledge entities into the QA pairs through the conceptualized augmentation of the entities (Wang et al., 2023) with the pre-defined natural language templates (e.g., the relation of *xWant* is transformed to *As a result, PersonX wanted to*). This process results in textual synthetic QA pairs (Q, A).

Synthetic VQA On the textual synthetic QA pairs, we input the textual question Q to the text-to-image model $\mathcal{M}_{T2I}$ to generate the visual counterpart I that depicts the scenarios described in each question. These generated images provide an additional layer of information, offering a visual context that enhances the reasoning ability based not only on textual descriptions but also on visual evidence. This augmentation leverages the strengths of visual imagery on language understanding (Gambrell and Bales, 1986; Dessalegn and Landau, 2013), potentially improving the robustness and accuracy of the model predictions.

However, relying solely on the synthetic relationships between QA pairs and generated images can introduce challenges related to the alignment of visual content since machines often fail to generate well-aligned images with textual inputs (Feng et al., 2023). Therefore, we augment the synthetic VQA pairs with the widely used Visual Commonsense Reasoning (VCR) dataset (Zellers et al., 2019). Each pair from this dataset consists of (Q, A, R, I), where R is a rationale for the correct answer; however, we omit R since our focus is on the QA pairs associated with relevant images. Additionally, to enrich the input and enhance visual comprehension for PLMs, we generate textual context information for each image using an image captioning model[2], which we prepend as a prefix to each Q[3].

3.3 Pre-training IMAGINE on Synthetic VQA

Based on the synthetic VQA dataset, we integrate commonsense knowledge into the models. Since IMAGINE involves two distinct modalities (i.e., text and image), we introduce two separate objectives to select the best answer choice: Language Modeling (LM) and Image-Text Matching (ITM). To obtain the LM scores, we calculate the masked language modeling loss for the Transformer encoder-based model, formulated as:

$$S_{LM}(T) = -\frac{1}{m}\sum_{t=1}^{m} \log P(w_t|...w_{t-1}, w_{t+1}...). \tag{1}$$

For the decoder-based model, we compute the autoregressive language modeling loss, defined as:

$$S_{LM}(T) = -\frac{1}{m}\sum_{t=1}^{m} \log P(w_t|w_1...w_{t-1}), \tag{2}$$

where w_i denotes the i-th word, and m is the number of tokens in the sequence T. To compute the ITM scores, we first contextualize the visual features based on the textual sequences. Let the visual features from the visual encoder $\mathcal{M}_I$ be denoted as

[2]We use InstructBLIP (Dai et al., 2023) for captioning.
[3]More details of synthetic VQA are in Appendix A.

V, we derive the contextualized visual features as follows:

$$C = \text{softmax}(\frac{\bar{T}V^\top}{\sqrt{d_v}})V, \quad (3)$$

where $\bar{T}$ is the feature vector from the PLMs $\mathcal{M}_T$. For the encoder-based model, we use the final hidden state of the [CLS] token as the context vector, and for the decoder-based model, we use the hidden state of the last token as the context vector. d_v is the dimension of visual features. We then achieve the ITM scores by calculating the similarity between contextualized visual features and textual features as follows:

$$S_I(T, V) = \text{sim}(\vec{T}, C), \quad (4)$$

where $\text{sim}(\cdot)$ denotes the cosine similarity function. By combining two different scores, we produce the joint scores S_J as follows:

$$S_J(T, V) = \frac{1}{2}(S_M(T) + S_I(T, V)), \quad (5)$$

After calculating all scores $S^{(1)}, S^{(2)}, ..., S^{(n)}$ for n answer candidates, we calculate the marginal ranking loss defined as:

$$\mathcal{L}_{QA}(S) = \frac{1}{n}\sum_{i=1, i\neq y}^{n} \max(0, \eta - S^{(y)} + S^{(i)}), \quad (6)$$

where y indicates the index of the correct answer and η is the pre-defined margin. The overall objectives are as follows:

$$\mathcal{L} = \mathcal{L}_{QA}(S_M) + \mathcal{L}_{QA}(S_I) + \mathcal{L}_{QA}(S_J). \quad (7)$$

However, we have empirically observed that the ITM objective prevents the model from learning the LM objective, which is essential for developing reasoning capabilities. To mitigate the conflict between these two objectives, we introduce two distinct adapters (He et al., 2022a), LM adapter and ITM adapter. Each adapter is trained separately with a different focus. It is important to note that only the weights within these adapters are optimized during training; all other parameters remain frozen. By separating the parameters for objectives, we can effectively reduce conflicts between them.

3.4 Inference from IMAGINE

For the zero-shot evaluation, we use the same strategy to compute the LM and ITM scores after synthesizing the image based on the question. However, we ensemble two scores to derive the model's prediction after obtaining the probability distribution through softmax.

$$P(S) = \text{softmax}(S^{(1)}, S^{(2)}, ..., S^{(n)}), \quad (8)$$

$$P(A|Q) = (1 - \lambda) \cdot P(S_M) + \lambda \cdot P(S_I), \quad (9)$$

where λ is an ensemble coefficient that controls the contributions between textual and visual features.

4 Experiments

In this section, we demonstrate the effectiveness of IMAGINE. Specifically, we conduct extensive experiments and analysis to answer the following research questions:

Q1 (Generalizability) Does IMAGINE offer better zero-shot performance across a broad range of reasoning benchmarks? (§4.2)

Q2 (Multimodality) Does IMAGINE effectively integrate visual signals (imagination) with textual knowledge? (§4.3, §4.4)

Q3 (Effectiveness) How effective are the components of IMAGINE in zero-shot commonsense reasoning? (§4.5)

4.1 Experimental Setup

Dataset. Following the previous works on zero-shot reasoning (Ma et al., 2021; Yang et al., 2022), we evaluate our framework on commonsense reasoning tasks and science QA tasks to assess its generalizability. Specifically, we evaluate each baseline on the five reasoning benchmarks, including Abductive NLI (αNLI; Bhagavatula et al., 2020), CommonsenseQA (CSQA; Talmor et al., 2019), PhysicalIQA (PIQA; Bisk et al., 2020), SocialIQA (SIQA; Sap et al., 2019b), and Winogrande (WG; Sakaguchi et al., 2020). These datasets vary significantly in format (e.g., natural language inference, QA, pronoun resolution) and required knowledge (e.g., social and physical knowledge for SIQA and PIQA, respectively), enabling a comprehensive evaluation of a wide spectrum of reasoning capabilities. For science QA tasks, we assess each baseline on the four benchmarks, including QA via Sentence Composition (QASC; Khot et al., 2020), Science Questions (SciQ; Welbl et al., 2017), and the AI2 Reasoning Challenge (ARC-Easy, ARC-Challenge; Clark et al., 2018). Given that science QA datasets often contain various types of reporting bias, such as color and shape biases, we selected these datasets to verify the efficacy of IMAGINE in mitigating reporting bias.

Method	KB	αNLI	CSQA	PIQA	SIQA	WG	Avg.
GPT-2-L (Radford et al., 2019)	-	56.5	41.4	68.9	44.6	53.2	52.9
RoBERTa-L (Liu et al., 2019)	-	65.6	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L (He et al., 2023)	-	59.9	25.4	44.8	47.8	50.3	45.6
RoBERTa-L (MR; Ma et al., 2021)	AT	70.8	64.2	72.1	63.1	59.6	66.0
Zero-shot Fusion (Kim et al., 2022)	AT, CN, WD, WN	72.5	68.2	72.9	**66.6**	60.8	68.2
CAR-RoBERTa-L (Wang et al., 2023)	AbsAT	72.7	66.3	73.2	64.0	62.0	67.6
CAR-DeBERTa-v3-L (Wang et al., 2023)	AbsAT	79.6	69.3	78.6	64.0	78.2	73.9
CANDLE-DeBERTa-v3-L (Wang et al., 2024)	CANDLE	81.2	69.9	80.3	65.9	**78.3**	75.1
CANDLE-VERA-T5-xxl (Wang et al., 2024)	CANDLE	73.8	64.7	77.6	59.4	71.3	69.4
IMAGINE-GPT-2-L	Synthetic VQA	61.5	63.9	68.9	53.0	55.2	58.5
IMAGINE-RoBERTa-L	Synthetic VQA	74.7	67.5	72.3	64.3	61.2	68.0
IMAGINE-DeBERTa-v3-L	Synthetic VQA	**82.2**	**74.0**	**80.7**	66.3	76.7	**76.0**
Human	-	91.4	88.9	94.9	86.9	94.1	91.2

Table 1: Zero-shot evaluation results on commonsense reasoning tasks (Accuracy %). **Bold** and Underline indicate the best and second-best results, respectively. AT, CN, WD, WN, and AbsAT refer to ATOMIC, ConcetNet, WikiData, WordNet, and AbstractATOMIC. The full comparison is presented in Table 13 (Appendix). The results are from each reference.

Method	αNLI	CSQA	PIQA	SIQA	WG	Avg.
GPT-3.5	61.8	68.9	67.8	68.0	60.7	65.4
ChatGPT	73.2	**75.7**	81.7	**69.7**	64.1	72.9
GPT-4	75.0	43.0	73.0	57.0	**77.0**	65.0
LLaMA2$_{13B}$	55.9	67.3	80.2	50.3	72.8	65.3
Mistral$_{7B}$	51.0	59.6	**83.0**	42.9	75.3	62.4
IMAGINE	**82.2**	74.0	80.7	66.3	76.7	**76.0**
Human	91.4	88.9	94.9	86.9	94.1	91.2

Table 2: Zero-shot evaluation results of LLMs on commonsense reasoning tasks (Accuracy %). **Bold** and Underline indicate the best and second-best results, respectively. Results are taken from Wang et al. (2024), and IMAGINE represents the results on DeBERTa-v3-L.

Method	QASC	SciQ	ARC-E	ARC-C
SMLM*	26.6	-	33.4	28.4
CAR-RoBERTa-L	56.7	60.7	57.0	36.5
CAR-DeBERTa-v3-L	70.0	76.9	75.3	53.2
OPT$_{30B}$*	39.7	72.7	58.2	34.8
FLAN$_{137B}$*	-	-	**79.5**	**61.7**
Z-LaVI (RoBERTa-L)*	27.2	51.3	51.8	33.4
Z-LaVI (BART-L)*	27.3	51.0	56.1	36.5
Z-LaVI (OPT$_{30B}$)*	42.1	74.0	59.5	34.1
IMAGINE-GPT-2-L	46.5	58.4	55.1	35.1
IMAGINE-RoBERTa-L	57.1	63.7	57.9	39.1
IMAGINE-DeBERTa-v3-L	**72.4**	**78.9**	76.0	56.2

Table 3: Zero-shot evaluation results on four science question-answering tasks (Accuracy %). **Bold** and Underline indicate the best and second-best results, respectively. Results (*) are taken from references (Banerjee and Baral, 2020; Yang et al., 2022; Wei et al., 2022)

Baselines. We mainly compare IMAGINE with the following zero-shot commonsense reasoning frameworks: MR (Ma et al., 2021), SMLM (Banerjee and Baral, 2020), Zero-shot Fusion (Kim et al., 2022), CAR (Wang et al., 2023), and the state-of-the-art framework, CANDLE (Wang et al., 2024). To confirm the efficacy of training with machine imagination in IMAGINE, we also compare it with Z-LaVI (Yang et al., 2022), which leverages machine imagination but does not include the training process. Beyond the reasoning framework based on KBs, we evaluate the recent LLMs, which include LLaMA2$_{13B}$ (Touvron et al., 2023), Mistral$_{7B}$ (v0.1) (Jiang et al., 2023), OPT$_{30B}$ (Zhang et al., 2022), FLAN$_{137B}$ (Wei et al., 2022), and the GPT families (i.e., GPT-3.5, ChatGPT (`gpt-3.5-turbo`), GPT-4).

Backbones. To verify the general applicability of IMAGINE, we apply our method to the both encoder and decoder models. Specifically, following the previous works, we utilize RoBERTa-Large (Liu et al., 2019) and DeBERTa-v3-Large (He et al., 2023). Each model has 362M and 443M parameters, respectively. As for the decoder model, we use GPT-2-Large that involves 792M parameters. Implementation details are in Appendix B.

4.2 Main Results

Tables 1, 2, and 3 show the results for the commonsense reasoning tasks and the science question-answering tasks. Models based on IMAGINE reveal either superior or competitive performance on overall reasoning tasks. This demonstrates the effectiveness of IMAGINE and highlights the benefit of leveraging machine imagination for reasoning.

In particular, compared to zero-shot common-

[PIQA]
Q: Brush dust off eyebrows
A1: Use toothbrush to groom
A2: Use dental floss to groom

Existing Model: A2(X) IMAGINE: A1(O)

[aNLI]
Q. Everyone laughed at the funny video.
A1. They took a study break to film videos
A2. Beth found a funny cat video.

Existing Model: A1(X) IMAGINE: A2(O)

[CSQA]
Q. What part of a table would you put a ruler in?
A1. Drawer A2. Desk
A3. The backside A4. Office
A5. Measure distance

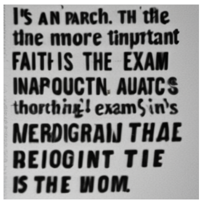
Existing Model: A3(X) IMAGINE: A3(X)

[SIQA]
Q. After starting the bar fight Kai told Riley that he had better go quickly. How would you describe Riley?
A1. A trouble maker A2. Full of adrenaline
A3. A peace maker

Existing Model: A1(O) IMAGINE: A1(O)

[CSQA]
Q. Where usually lacks an elevator but sometimes has a telephone book?
A1. At hotel A2. Kitchen
A3. Library **A4. Telephone booth**
A5. House

Existing Model: A4(O) IMAGINE: A5(X)



[WG]
Q. It is an article of faith that the paper is more important than the exam, even though the _ weighs less heavily on the grade.
A1. Paper A2. Exam

Existing Model: A2(X) IMAGINE: A2(X)

Figure 4: Comparison of model predictions and the correctness from IMAGINE and the existing model (Wang et al., 2023) on five commonsense reasoning tasks.

sense reasoning frameworks in commonsense reasoning tasks (Table 1), IMAGINE-DeBERTa-v3-L model surpasses the previous state-of-the-art by 0.9%p on average, and specifically by 4.1%p on the CSQA. This suggests that synthetic VQA significantly enhances generalization performance in zero-shot commonsense reasoning. Comparison results with LLMs (Table 2) also shows that IMAGINE outperforms recent LLMs, including ChatGPT and GPT-4 (OpenAI, 2023). This result suggests the superior efficiency and effectiveness of IMAGINE's multimodal approach.

IMAGINE also proves effective for science QA tasks (Table 3). Compared to the models with KBs and larger models, IMAGINE presents better or competitive reasoning performance. These results confirm the effectiveness of the machine imagination capabilities on science-related contexts. We also highlight the comparison results with Z-LaVI (Yang et al., 2022) that leverages imagination similar to ours. IMAGINE outperforms this method by a significant margin (18.5%p on average), underscoring the importance of the pre-training phase in effectively utilizing machine imagination.

4.3 Impact of Imagination on Model Inference

We analyze the inference results from the text-based model, CAR (Wang et al., 2023), and IMAGINE to confirm the impact of machine imagination on the model inference. The results are shown in Figure 4. We draw three major findings regarding the impact of imagination: (i) When the text contains limited commonsense knowledge, imagination indeed helps the model to correctly infer the answer (First row in the Figure), i.e., positive impact on predictions (ii) When the generated images only partially capture the context of the text query, imagination does not affect the inference results (Second row in the Figure). (iii) When images deviate from the real world, imagination can lead to incorrect inferences (Third row in the Figure). Specifically, we empirically observe that longer text queries often result in such cases. These results suggest that incorporating a text-to-image model with better alignment capabilities could potentially mitigate the negative impacts of imagination[4].

KB	αNLI	CSQA	PIQA	SIQA	WQ	Avg.
Synthetic VQA	<u>74.7</u>	<u>67.5</u>	<u>72.3</u>	<u>64.3</u>	**61.2**	**68.0**
w/o VCR	71.7	65.7	**72.3**	**65.7**	<u>60.3</u>	<u>67.1</u>
w/o AbsAT	**75.6**	**67.5**	71.7	56.2	58.8	66.0
w/o VCR, AbsAT	65.6	45.0	67.6	47.3	57.5	56.6

Table 4: Ablation results on Synthetic VQA. **Bold** and <u>underline</u> indicate the best and second-best results.

4.4 Contributions of Synthetic VQA

To confirm the effectiveness of each component in Synthetic VQA, we evaluate the contribution of

[4]We provide more examples with the visualization of model attention in Appendix F.

LM	ITM	αNLI	CSQA	PIQA	SIQA	WG	Avg.
✓	✓	**74.7**	**67.5**	**72.3**	**64.3**	**61.2**	**68.0**
✓	-	74.3	65.2	71.9	62.3	60.5	66.8
-	✓	71.7	62.0	68.8	60.0	59.6	64.4
-	-	65.6	45.0	67.6	47.3	57.5	56.6

Table 5: Ablation results on pre-training objective of IMAGINE. We use a RoBERTa-L as a backbone.

Inference	αNLI	CSQA	PIQA	SIQA	WG	Avg.
Ensemble	**74.7**	**67.5**	**72.3**	**64.3**	**61.2**	**68.0**
LM	74.1	66.9	71.8	63.8	61.1	67.1
ITM	71.7	63.1	68.3	59.8	59.4	64.0

Table 6: Results of the different inference strategy (LM, ITM). These strategies are evaluated on RoBERTa-L.

Model	αNLI	CSQA	PIQA	SIQA	WG	Avg.
Adapter	**74.7**	**67.5**	**72.3**	**64.3**	**61.2**	**68.0**
Full	73.0	65.4	71.1	61.5	61.2	66.4

Table 7: Evaluation results of IMAGINE with full fine-tuning (Full) and adapter tuning (Adapter).

AbsAT and VCR. Table 4 presents the results on commonsense reasoning tasks. The model trained only with AbsAT (i.e., w/o VCR) shows superior performance on datasets that contain longer sequences and require complex knowledge (e.g., PIQA, SIQA). In contrast, the model trained only with VCR (i.e., w/o AbsAT) shows its strength on the dataset that contain simpler questions (αNLI, CSQA) which allows the better use of visual information. When combining these two components, the synthetic VQA results in well-generalized reasoners across diverse reasoning tasks, demonstrating the complementary effect of each component.

4.5 Component Analysis on IMAGINE

Ablation on Training Objectives. IMAGINE employs two objectives (i.e., LM, ITM) to learn commonsense knowledge from different modalities. We perform ablations on these objectives to verify their contributions in enhancing zero-shot reasoning capabilities. Table 5 shows the ablation results. Notably, omitting the LM objective leads to a significant drop in performance, underscoring the crucial role of language understanding in commonsense reasoning. Furthermore, while ITM alone does not significantly impact reasoning effectiveness, combining ITM with LM results in improved reasoning performance. These findings suggest that integrating visual information in model optimization leads to better reasoning in commonsense situations.

Effect of Ensemble Inference. IMAGINE performs reasoning based on ensemble of the LM and ITM scores. To investigate the contributions in scores obtained from these two different modalities, we evaluate each score independently. The results are presented in Table 6. We observe the lowest performance when evaluating only the ITM scores. However, ensembling LM scores with the ITM results in significant performance improvement across all tasks, even though the scores derived from images are much lower than those from text. This indicates that integrating machine-generated images can complement and enhance language-based reasoning abilities[5].

Impact of Adapter. IMAGINE utilizes adapters (He et al., 2022a) to alleviate the conflicts between the two objectives (i.e., LM, ITM) during the pre-training. In this study, we examine whether separating parameters through adapters for distinct modality objectives is truly effective. Table 7 presents the ablation results on adapters. We observe a significant decline in reasoning performance when adapters are removed. This suggests that direct training of PLMs with images adversely affects the acquisition of textual knowledge. One plausible explanation for this phenomenon is possibly related to catastrophic forgetting (Kirkpatrick et al., 2017), where the model loses previously acquired knowledge (i.e., textual knowledge inherent in PLMs). This highlights the effectiveness of adapters in maintaining the model's linguistic understanding when it learns from new modalities.

5 Conclusion

In this paper, we have proposed IMAGINE, a novel zero-shot commonsense reasoning framework that leverages visual signals to mitigate reporting bias. To steer IMAGINE in effectively utilizing visual information, we have created a large-scale synthetic VQA dataset and optimized the model to jointly use both textual and visual information. We have conducted extensive experiments across a broad range of reasoning tasks. Comprehensive results have shown that IMAGINE establishes new state-of-the-art results on zero-shot commonsense reasoning tasks compared to strong baselines, demonstrating the efficacy of machine imagination.

[5] More analysis on ensemble methods are in Appendix D.

Limitations

We have demonstrated the efficacy of the machine imagination to improve zero-shot commonsense reasoning ability. However, we still have the following limitations:

Additional Computations While machine imagination leads to performance improvement in PLMs, it necessitates additional computations for generating and processing visual signals. This limitation can be addressed by retrieving relevant images instead of synthesizing new ones, as demonstrated in previous work (Yang et al., 2022). We consider this approach a promising avenue for future research.

Exploration of IMAGINE on LLMs In this work, we apply IMAGINE to only intermediate-size models (300M to 790M), as one of our objectives is to show the smaller models with machine imagination outperforms LLMs on a broad range of commonsense reasoning tasks. However, we believe that IMAGINE can be effectively combined with LLMs, given that the reporting bias is an inherent issue in the pre-training corpus and not the models themselves. We plan to explore the scaling of machine imagination in our future research.

References

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, The Eleventh Symposium on Educational Advances in Artificial Intelligence*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*.

Banchiamlack Dessalegn and Barbara Landau. 2013. Interaction between language and vision: It's momentary, abstract, and it develops. *Cognition*, 127:331–344.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Zi-Yi Dou and Nanyun Peng. 2022. Zero-shot commonsense question answering with cloze translation and consistency optimization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, The Twelveth Symposium on Educational Advances in Artificial Intelligence*.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh*

International Conference on Learning Representations.

Linda B. Gambrell and Ruby J. Bales. 1986. Mental imagery and the comprehension-monitoring performance of fourth- and fifth-grade poor readers. *Reading Research Quarterly*, 21:454.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*.

Xin Guan, Biwei Cao, Qingqing Gao, Zheng Yin, Bo Liu, and Jiuxin Cao. 2023. Multi-hop commonsense knowledge injection framework for zero-shot commonsense question answering. *CoRR*, abs/2305.05936.

David Gunning. 2018. Machine common sense concept paper. *arXiv preprint arXiv:1810.07528*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022a. Towards a unified view of parameter-efficient transfer learning. In *The Tenth International Conference on Learning Representations*.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022b. Acquiring and modelling abstract commonsense knowledge via conceptualization. *CoRR*, abs/2206.01532.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Yu Jin Kim, Beong-woo Kwak, Youngwook Kim, Reinald Kim Amplayo, Seung-won Hwang, and Jinyoung Yeo. 2022. Modularized transfer learning with multiple knowledge graphs for zero-shot commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 19730–19742.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yujie Lu, Wanrong Zhu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-augmented natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *CoRR*, abs/1909.08855.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, The Thirty-First Innovative Applications of Artificial Intelligence Conference, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Ying Su, Zihao Wang, Tianqing Fang, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. MICO: A multi-alternative contrastive learning framework for commonsense knowledge representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

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.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, Chunkit Chan, and Yangqiu Song. 2024. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. *CoRR*, abs/2401.07286.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth*

International Conference on Learning Representations.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*.

Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. Z-lavi: Zero-shot language solver fueled by visual imagination. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

Appendix

A Synthetic VQA dataset

	Train	Dev	Total
# Images generated from AbsAT	18,838	1,695	20,533
# QA pairs from AbsAT	486,778	46,238	533,016
# Images from VCR	80,418	9,929	90,347
# QA pairs from VCR	212,923	26,534	239,457
# Total Images	99,256	11,624	110,880
# Total QA pairs	699,701	72,772	772,473

Table 8: Statistic of synthetic VQA dataset.

We construct a synthetic VQA dataset using AbstractATOMIC and VCR. First, we generate images using the questions from AbstractATOMIC. Since AbstractATOMIC consists only of text, we need to create images based on these questions. In this process, we standardize all the person names in the questions to "Person" and remove duplicate questions, resulting in approximately 20K images. To include more realistic images and commonsense questions corresponding to those images, we extract question-answer pairs from VCR images. However, most of these questions are directly related to the images, making it difficult to answer without them, which poses a challenge for LM-based training. To address this, we replace the person indices in the questions with gender-neutral names and generate captions for the images to use as prefixes for the questions. In addition, each QA pair from VCR has four answer candidates, while each pair from AbstractATOMIC has three candidates. To combine them, we match the number of answer choices by randomly discarding one distractor from VCR. The statistic of our dataset is provided in Table 8.

B Implementation Details

To construct the VQA pairs, we primarily use DALL-E 3-XL (Betker et al., 2023), a powerful image synthesis model. For generating images in the synthetic VQA dataset, we first remove overly specific information, such as personal names, from the questions. Then, we generate images with a resolution of 384 × 384 using 50 inference steps. During the evaluation, we generate 512 × 512 images for each task based on the questions, maintaining the same number of inference steps. We use the CLIP-Large (Radford et al., 2021) model to extract image features. Following prior work, we use two powerful PLMs as the backbone. We add Parallel Adapter (He et al., 2022a) with a reduction factor of 16 to each model and freeze all parameters except for the adapters. We follow the training settings of Ma et al. (2021) and Wang et al. (2023) to train Transformer decoder-based and encoder-based model for the in-depth comparison. We report our results derived from the ensemble score using the optimal ensemble weight for each task. All experiments are conducted using four NVIDIA A5000 GPUs. More details are presented in Table 9.

IMAGINE	GPT-2-L	RoBERTa-L	DeBERTa-v3-L
Image Encoder		CLIP-ViT-L/14	
# Params.	792M + 428M	362M + 428M	443M + 428M
# Trainable Params.	7.9M	8.4M	8.4M
Training Time	70h	30h	80h
Batch Size		8, 16, **32**, 64	
Learning Rate		7e-6, **1e-5**, 3e-5	
Epoch		2	

Table 9: Detailed training settings for IMAGINE. **Bold** indicates the chosen hyperparameter.

C Impact of Image Quality

We aim to observe the changes in inference performance based on image quality by generating images of various qualities using three different methods. First, similar to our main experiment, we utilize the questions from the evaluation dataset to generate images with a resolution of 512 × 512 using both DALL-E 3-XL and the Latent Diffusion Model (LDM; Rombach et al., 2022), which has relatively lower image synthesis capabilities. Additionally, we generate images with a resolution of 384 × 384 using DALL-E 3-XL, following the same method used for creating the synthetic VQA dataset.

IMAGINE	αNLI	CSQA	PIQA	SIQA	WG	Avg.
Text only	73.2	66.3	71.3	64.5	60.3	67.1
LDM (512 × 512)	73.2	66.3	71.9	64.3	60.6	67.3
DALL-E 3 (384 × 384)	74.5	66.8	71.9	64.3	60.6	67.6
DALL-E 3 (512 × 512)	**74.7**	**67.5**	**72.3**	**64.3**	**61.2**	**68.0**

Table 10: Results of using various image synthesis models for evaluation. The numbers in parentheses indicate the image resolution.

The results in Table 10 show that the IMAGINE with the LDM model performs the worst, indicating that utilizing a less effective image synthesis model can degrade overall performance. However, all models benefit from incorporating various resolutions of images. As seen in Figure 5, this is likely because the generated images, despite varying in

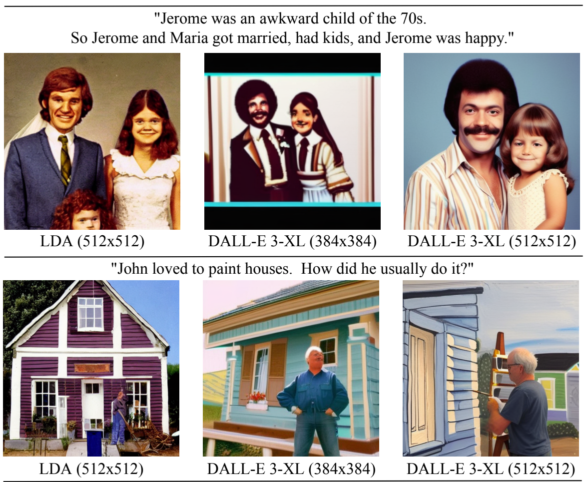


Figure 5: Comparison of generated images. The sentences are the queries used to generate the images.

quality, mostly maintain contextual relevance to the query sentences, thereby having a similar positive impact on the inference results.

D Ensemble Methods

To verify the effectiveness of our framework's multimodality approach, we train two unimodal models using different seeds on the synthetic VQA dataset, utilizing only the text. We then ensemble the scores obtained from these two models. The results are presented in Table 11. While ensembling scores from single modalities (LM+LM) provides performance benefits, ensembling scores from two different modalities (LM+ITM), as done in IMAGINE, proves to be the most effective. This demonstrates that the multimodality approach plays a crucial role in enhancing zero-shot reasoning performance.

RoBERTa-Large	αNLI	CSQA	PIQA	SIQA	WG	Avg.
LM	74.3	65.2	71.9	62.3	60.5	66.8
LM+LM	74.3	66.0	72.1	64.2	60.4	67.4
LM+ITM (IMAGINE)	**74.7**	**67.5**	**72.3**	**64.3**	**61.2**	**68.0**

Table 11: Results of two different ensemble methods.

We report the optimal ensemble weights used for our framework in Figure 6. The larger the ensemble weight, the greater the influence of the image scores. Additionally, we draw a line indicating the average accuracy in each plot. From this, we can infer that the DeBERTa-v3-Large model utilizes image information more extensively than the RoBERTa-Large. When applying IMAGINE to DeBERTa-v3-Large, the performance improvement is greater than when using RoBERTa-Large, suggesting that visual information contributes positively to most reasoning tasks.

E IMAGINE with Decoder-based Model

We conducted experiments using GPT-2, a widely-used decoder-based generative language model, to verify the applicability to recent language models. We follow the settings of (Ma et al., 2021) to train to model on synthetic datasets.

	αNLI	CSQA	PIQA	SIQA	WG	Avg.
GPT-2-L	56.5	41.4	68.9	44.6	53.2	52.9
GPT-2-L (MR)	59.2	48.0	67.5	53.6	54.7	56.6
CAR-GPT-2-L	61.7	50.0	68.2	52.3	55.2	57.5
IMAGINE-GPT-2-L	61.5	53.9	68.9	53.0	55.2	58.5

Table 12: Zero-shot evaluation results with decoder-only generative model.

The results in Table 12 demonstrate that IMAGINE is effective not only for encoder-based models but also for decoder-based models. Based on these findings, we plan to address methodologies in future work that can effectively utilize images while preserving the rich language understanding capabilities of large language models.

F Visualization of Image Attention

We aim to visualize how the model utilizes specific parts of an image. The formula to compute contextualized visual features used for computing the ITM score calculation process is similar to the attention algorithm, allowing us to derive attention scores for each image patch. Based on these scores, we erase 100 image patches with the lowest scores to understand which parts the model focuses on. As shown in Figure 7, 8, and 9, each model tends to assign relatively high attention scores to objects related to the question in most cases, rather than using the image patches randomly. This is notable because the model can effectively capture the relationship between text and images using adapters, despite training with much less data compared to existing visual-language modeling studies (Li et al., 2023; Zhu et al., 2023). In addition, we observe that the DeBERTa-v3-Large model tends to focus more frequently on the correct parts than the RoBERTa-Large model. Figure 7 shows these cases clearly. This aligns with the result that the IMAGINE is more effective with DeBERTa-v3-Large, suggesting that a model with high generalization performance is also useful for learning new modalities.

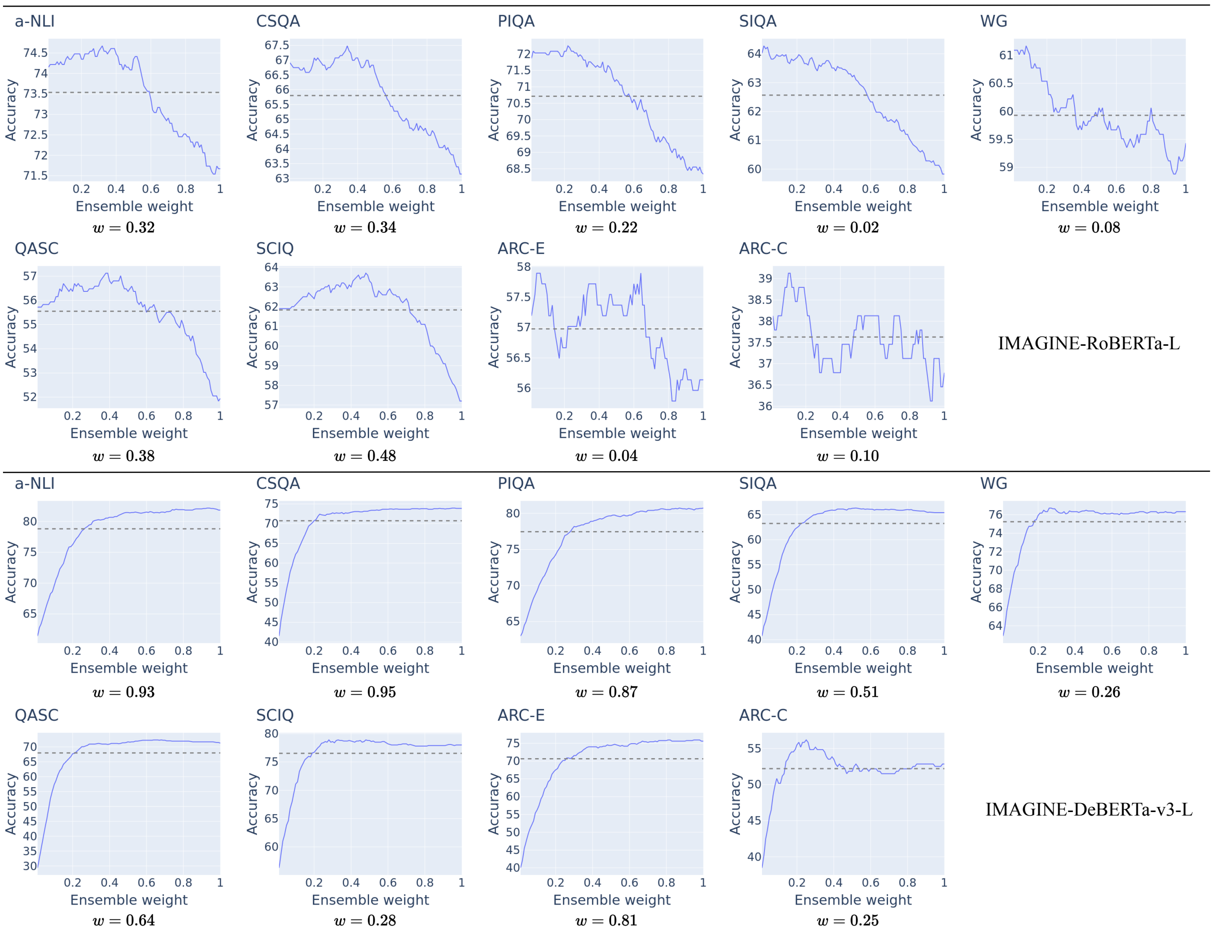


Figure 6: Model accuracy variation with different ensemble weights. The optimal w for each task is shown below the plots. The line in the middle indicates the average accuracy.

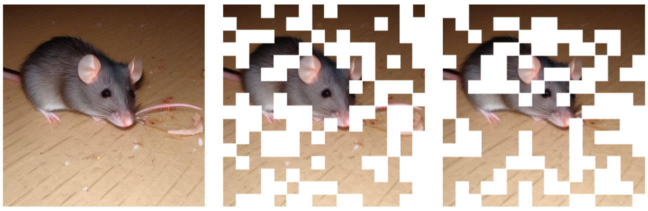
DeBERTa-v3-L | RoBERTa-L

Q. Joe was walking through downtown.
He reluctantly agreed to give them an interview.
A1. He was approached by a pretty woman.
A2. He was approached by a survey taker.

DeBERTa-v3-L: A2 (O) RoBERTa-L: A1 (X)

Q. I got up from a nap feeling very hungry.
After the inspector arrived and killed the rats, I felt very happy.
A1. I decided not to eat when I saw a rat in the kitchen.
A2. I ate a lot of rats in my kitchen.

DeBERTa-v3-L: A1 (O) RoBERTa-L: A1 (O)



Q. John didn't mind getting in line. It was what game after that he hated. The time, the sore feet. He did not like doing what?
A1. Have to wait for **A2. Standing in line** A3. Eat cake
A4. Less confusion A5. Being ordered

DeBERTa-v3-L: A2 (O) RoBERTa-L: A2 (O)

Q. Of all the sports, Billy enjoys football,
but what does his concerned mother think of the sport?
A1. Very entertaining A2. Fun A3. Competitive
A4. Competitive **A5. Violent**

DeBERTa-v3-L: A5 (O) RoBERTa-L: A2 (X)

Q. How to quickly cool down a bottled water drink?
A1. Run the paper towel under some water and wrap a bottle around it then place in the freezer for 20 minutes.
A2. Run the bottle under some water and wrap a paper towel around it then place in the freezer for 20 minutes.

DeBERTa-v3-L: A2 (O) RoBERTa-L: A2 (O)

Q. What is the best way to apply nail polish to a professional result?
A1. A quick way to apply nail polish is to use a large brush, then cover any messy areas with flesh-colored nail polish.
A2. Tape the cuticles with snugly fitting tape, then paint the nails. Remove the tape and use a nail polish remover-soaked q-tip to clean any excess polish from the cuticles or fingers.

DeBERTa-v3-L: A2 (O) RoBERTa-L: A2 (O)

Figure 7: Randomly sampled examples from IMAGINE alongside the visualization of image attention from the Abductive NLI, CommonsenseQA, and PIQA validation sets.

DeBERTa-v3-L RoBERTa-L

Q. Robin studied hard the night before, and found the test to be very easy. Robin finished the test quickly. How would Robin feel afterwards?
A1. Proud A2. Motivated
A3. Nervous

DeBERTa-v3-L: A1 (O) RoBERTa-L: A1 (O)

Q. Alex bought his entire team gold watches and when he gave them the present he put each watch on their wrist himself. How would you describe Alex?
A1. A greedy person A2. Satisfied over the gift he gave his team
A3. A thoughtful person

DeBERTa-v3-L: A1 (X) RoBERTa-L: A3 (O)

Q. As a parent, Catherine doesn't let her kids watch movies, but they can watch some TV chows. Catherine thinks the ___ are too violent.
A1. Movies A2. TV shows

DeBERTa-v3-L: A1 (O) RoBERTa-L: A1 (O)

Q. The farmer had more corn to harvest than yams because his cow hated eating the ___.
A1. Yam **A2. Corn**

DeBERTa-v3-L: A1 (X) RoBERTa-L: A1 (X)

Q. What cycle is the most directly affected by the combustion of fossil fuels?
A1. Rock cycle A2. Water cycle **A3. Carbon cycle**
A4. Nitrogen cycle

DeBERTa-v3-L: A3 (O) RoBERTa-L: A3 (O)

Q. What energy change takes place when a piece of bread is toasted in a toaster?
A1. Chemical energy to light energy
A2. Electrical energy to heat energy
A3. Heat energy to chemical energy
A4. Light energy to electrical energy

DeBERTa-v3-L: A2 (O) RoBERTa-L: A3 (X)

Figure 8: Randomly sampled examples from IMAGINE alongside the visualization of image attention from the SIQA, Winogrande, and ARC-easy validation sets.

DeBERTa-v3-L RoBERTa-L

Q. Where would it be most dangerous to work with electric tools?
A1. In a garage
A2. Beside a swimming pool
A3. Near a television or computer
A4. In a cool basement

DeBERTa-v3-L: A2 (X) RoBERTa-L: A2 (X)

Q. When the motion of liquid water molecules slow, what most likely happens?
A1. The liquid water forms a solid
A2. The liquid water condenses
A3. The liquid water undergoes a chemical change
A4. The liquid water becomes a vapor

DeBERTa-v3-L: A2 (X) RoBERTa-L: A3 (X)

Q. What is changing globally?
A1. The number of countries. A2. rapid growth
A3. How they move **A4. Temperature and moisture**
A5. Differences in speed A6. Net biomass
A7. Occurs over a wide range A8. Exposure to oxygen and water

DeBERTa-v3-L: A4 (O) RoBERTa-L: A1 (X)

Q. What has tiny hairs that trap particles?
A1. Sponges A2. Molecules
A3. Oaks A4. Lizards
A5. Protozoa A6. Snakes
A7. Cilia A8. Clouds

DeBERTa-v3-L: A7 (X) RoBERTa-L: A4 (X)

Q. What are the outer planets of the solar system made of?
A1. Solids A2. Plasma A3. Liquids **A4. Gases**

DeBERTa-v3-L: A4 (O) RoBERTa-L: A4 (O)

Q. What do we call cyclones that form in tropical latitudes?
A1. Eruptions A2. Twister A3. Disturbances **A4. hurricanes**

DeBERTa-v3-L: A4 (O) RoBERTa-L: A4 (O)

Figure 9: Randomly sampled examples from IMAGINE alongside the visualization of image attention from the ARC-challenge, QASC, and SciQ validation sets.

Method	KB	αNLI	CSQA	PIQA	SIQA	WG	Avg.
Pre-trained Language Models							
GPT-2-L (Radford et al., 2019)	-	56.5	41.4	68.9	44.6	53.2	52.9
RoBERTa-L (Liu et al., 2019)	-	65.6	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L (He et al., 2023)	-	59.9	25.4	44.8	47.8	50.3	45.6
Self-talk (Shwartz et al., 2020)	-	-	32.4	70.2	46.2	54.7	-
COMET-DynGen (Bosselut et al., 2021)	AT	-	-	-	50.1	-	-
SMLM (Banerjee and Baral, 2020)	*	65.3	38.8	-	48.5	*	-
GPT-2-L (MR; Ma et al., 2021)	AT	59.2	48.0	67.5	53.6	54.7	56.6
RoBERTa-L (MR; Ma et al., 2021)	AT	70.8	64.2	72.1	63.1	59.6	66.0
DeBERTa-v3-L (MR; Ma et al., 2021)	AT	76.0	67.0	78.0	62.1	76.0	71.8
MICO (Su et al., 2022)	AT	-	44.2	-	56.0	-	-
Zero-shot Fusion (Kim et al., 2022)	AT, CN, WD, WN	72.5	68.2	72.9	66.6	60.8	68.2
Multi-hop Knowledge Injection (Guan et al., 2023)	AT, CN, WD, WN	72.5	71.0	73.1	-	61.0	-
CAR-GPT-2-L (Wang et al., 2023)	AbsAT	61.7	50.0	68.2	52.3	55.2	57.5
CAR-RoBERTa-L (Wang et al., 2023)	AbsAT	72.7	66.3	73.2	64.0	62.0	67.6
CAR-DeBERTa-v3-L (Wang et al., 2023)	AbsAT	79.6	69.3	78.6	64.0	<u>78.2</u>	73.9
CANDLE-DeBERTa-v3-L (Wang et al., 2024)	CANDLE	81.2	69.9	80.3	65.9	**78.3**	<u>75.1</u>
Large Language Models							
GPT-3.5 (`text-davinci-003`)	-	61.8	68.9	67.8	<u>68.0</u>	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	-	73.2	**75.7**	81.7	**69.7**	64.1	72.9
GPT-4 (`gpt-4`)	-	75.0	43.0	73.0	57.0	77.0	65.0
LLAMA2-13B (Touvron et al., 2023)	-	55.9	67.3	80.2	50.3	72.8	65.3
Mistral-v0.1-7B (Jiang et al., 2023)	-	51.0	59.6	**83.0**	42.9	75.3	62.4
VERA-T5-xxl (Liu et al., 2023)	AT	71.2	61.7	76.4	58.2	67.2	66.9
VERA-T5-xxl (Liu et al., 2023)	AbsAT	73.2	63.0	77.2	58.1	68.1	68.0
CANDLE-VERA-T5-xxl (Wang et al., 2024)	CANDLE	73.8	64.7	77.6	59.4	71.3	69.4
Ours							
IMAGINE-GPT-2-L	Synthetic VQA	61.5	63.9	68.9	53.0	55.2	58.5
IMAGINE-RoBERTa-L	Synthetic VQA	74.7	67.5	72.3	64.3	61.2	68.0
IMAGINE-DeBERTa-v3-L	Synthetic VQA	**82.2**	<u>74.0</u>	<u>80.7</u>	66.3	76.7	**76.0**
Supervised & Human							
RoBERTa-L (Supervised)	-	85.6	78.5	79.2	76.6	79.3	79.8
DeBERTa-v3-L (Supervised)	-	89.0	82.1	84.5	80.1	84.1	84.0
Human	-	91.4	88.9	94.9	86.9	94.1	91.2

Table 13: Zero-shot evaluation results on five commonsense reasoning tasks (Accuracy %). **Bold** and <u>Underline</u> indicate the best and second-best results, respectively. AT, CN, WD, WN, and AbsAT refer to ATOMIC, ConcetNet, WikiData, WordNet, and AbstractATOMIC. The results of the large language models including GPT series are taken from Wang et al. (2024). SMLM (*) used different KBs for the different benchmarks.