COMPASS: Estimating COMmonsense Plausibility of Attributes Through Semantic Shifts

Anonymous ACL submission

Abstract

Accurately estimating commonsense plausibility is critical for ensuring the rationality of sentences generated by large language models. Inspired by the omission of commonsense in the text often does not affect the transmission of unnecessary information, in this work, we propose a novel method for estimating **COM**monsense **P**lausibility of **A**ttributes through **S**emantic **S**hifts (COMPASS). Instead of directly predicting attribute distributions, we focus on ranking attribute values based on their natural fit with a given object. COMPASS can be applied across both language models and multimodal language models (MLMs). Experimental results on the CoDa and ViComTe datasets show that our method significantly outperforms existing methods. The results also highlight the effectiveness of MLMs in learning attribute commonsense and show that integrating contrastive learning enhances sentence representations, which further improves performance. This work offers a novel perspective for broader commonsense plausibility estimation.

1 Introduction

Text generation in large language models (LLMs) continues to advance, yet ensuring that the generated outputs align with commonsense knowledge remains a challenge (Marcus, 2020; Elazar et al., 2021; Mahowald et al., 2024; Chen et al., 2023). Although several approaches have been proposed to enhance the commonsense consistency of generated text (Liu et al., 2021a; Wang et al., 2021; Feng et al., 2022; Wang et al., 2022b; Cui et al., 2024), the training and evaluation of these methods remains reliant on traditional metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2002), and METEOR (Banerjee and Lavie, 2005). These metrics focus on surface-level overlap with reference sentences, and can not assess whether a generated sentence adheres to commonsense knowledge. As a result, the development of metrics for evaluating commonsense plausibility is crucial for advancing techniques aimed at improving the quality of text generated by LLMs.

Assessing the rationality of commonsense is a challenging problem. Many prior works rely on the internal knowledge of language models (LMs) to estimate commonsense plausibility (Liu et al., 2023; Trinh and Le, 2018; Davison et al., 2019; Liu et al., 2021b; Zhao et al., 2024; Jung et al., 2022; Tafjord et al., 2022). In addition to the internal knowledge, some approaches augment LMs with external knowledge using retrieval augmentation (Zhang et al., 2021; Yu et al., 2022) or commonsense knowledge bases (KBs) (Choi et al., 2022; Ghazarian et al., 2023; Tian et al., 2023). Among them, some of the latest works (Ghazarian et al., 2023; Tian et al., 2023) simplify the key information in a sentence into triplets, avoiding the problem of information coupling. And they are applicable to specific types of commonsense, such as event knowledge.

In this work, we also highly simplify the sentences and focus on attribute commonsense of objects, such as color, shape, and material. An object may have many values for one attribute, such as the color of an apple can be red, green, etc. And the distribution data for each attribute value may not be completely accurate. Therefore, we focus on the attribute value ranking task. The model's performance is measured by its ability to rank more plausible attribute values higher.

Commonsense encompasses the fundamental knowledge and experiences shared by individuals in daily life (Schank, 1983; Winograd, 1986; Hobbs, 1990). From a linguistic perspective, commonsense often manifests implicitly in communication, the omission of commonsense in the text often does not affect the transmission of unnecessary information (Clark, 1996; Noveck and Sperber, 2004). Inspired by this, we propose the idea of measuring the **COM**monsense **P**lausibility of

Attributes through Semantic Shifts (COMPASS). That is, when a sentence lacks specific attributes, it tends to reflect an average representation of that object. However, when an attribute is explicitly introduced, the sentence undergoes a semantic shift. The degree of this shift can be used as an indicator of how natural or typical the attribute value is for that object. Therefore, it provides a meaningful way to rank the plausibility of different attribute values. For instance, when describing a 'penguin', the sentence 'there is a penguin' should undergo minimal semantic change when modified to 'there is a black penguin'. While a more implausible modification, such as 'there is a green penguin', would result in greater semantic shifts.

Our method builds upon pre-trained models that have been exposed to vast amounts of data, enabling them to learn a wide range of knowledge. In our framework, we consider incorporating both LMs and multimodal language models (MLMs). While LMs are widely used to assess knowledge, they often struggle with accurately modeling commonsense due to reporting bias in textual data (Gordon and Durme, 2013). On the other hand, attribute knowledge is inherently present in visual data. As illustrated in Figure 1a, MLMs learn to to align images and text, are suited to capturing this type of knowledge effectively. Moreover, our method relies on measuring semantic similarity, making both the quality of sentence representations and the model's ability to differentiate semantics critical. For models with weaker sentence representation capabilities, integrating contrastive learning techniques can substantially improve their effectiveness combined with our method.

We conduct extensive experiments on two datasets: the Color Dataset (Paik et al., 2021) and the Visual Commonsense Test (Zhang et al., 2022), which contains attribute distribution of objects on color, shape, and material. To evaluate the effectiveness of COMPASS, we compare its performance against existing commonsense plausibility estimation methods and test COMPASS across a wide range of backbones. Our experiments demonstrate the effectiveness of our method across various model architectures. First, COMPASS consistently improves performance across both LMs and MLMs, confirming its adaptability. Second, we find that MLMs, which incorporate both visual and textual information, achieve better effectiveness compared to LMs. This highlights the advantage of leveraging visual data in modeling attribute commonsense. Lastly, the quality of sentence representations is critical for the success of COMPASS. Models with weaker representation capabilities benefit significantly from contrastive learning techniques, which help refine their ability to represent sentences, further enhancing their performance in our framework.

Our contributions can be summarized as follows: (1) We propose a novel perspective for estimating the commonsense plausibility of attributes based on semantic shifts. (2) We conduct thorough experiments on two datasets, showing that COMPASS is effective across various models, including both multimodal and text-only architectures. (3) We show that MLMs are more efficient at learning attribute commonsense knowledge, and highlight the effectiveness of visual information for specific types of commonsense.

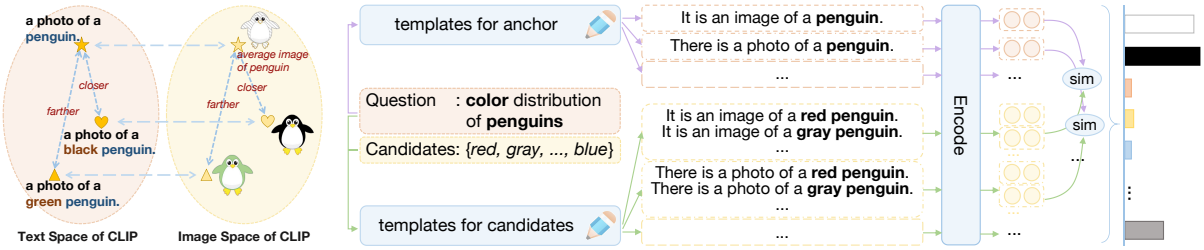
2 Related Work

2.1 Commonsense Plausibility Estimation Based on Internal Knowledge

The sentence probability and perplexity computed by LMs can serve as indicators of commonsense plausibility to a certain extent, even in zero-shot settings (Trinh and Le, 2018; Davison et al., 2019; Liu et al., 2021b). This is particularly evident in large LMs pre-trained on extensive corpora, where commonsense evaluation can be performed directly through prompts (Zhao et al., 2024). To further improve the performance, VERA (Liu et al., 2023) treats commonsense plausibility estimation as a binary classification task and trains a LM on ~7M commonsense statements. Beyond direct estimation, some methods (Jung et al., 2022; Tafjord et al., 2022) evaluate the plausibility of hypotheses by scoring the validity of entailment paths generated by the LMs.

2.2 Commonsense Plausibility Estimation Based on External Knowledge

The internal knowledge of LMs may be either insufficient or inaccurately modeled, prompting some methods to incorporate external knowledge to better estimate the rationality of commonsense. A typical approach is to supplement knowledge through retrieval sentences as augmentation (Zhang et al., 2021; Yu et al., 2022). Commonsense KBs, such as ConceptNet (Speer et al., 2016) and Atomic (Sap et al., 2019; Hwang et al., 2020), , store extensive commonsense knowledge. These resources



(a) (b)

Figure 1: (a) The text and image representation spaces of CLIP about 'penguin'. (b) How COMPASS works under multiple templates ensemble.

allow for the extraction of relevant subgraphs to assess sentence alignment with commonsense (Choi et al., 2022). However, KBs are inherently incomplete in their coverage. To address this limitation and harness the extensive knowledge embedded in LMs, COMET (Bosselut et al., 2019) introduced a dynamic KB by pre-training on existing commonsense KBs. Consequently, methods leveraging this dynamic KB (Ghazarian et al., 2023; Tian et al., 2023) demonstrate improved generalization across various commonsense tasks.

3 Attribute Value Ranking

In this work, we focus on object-attribute value ranking tasks to assess a model's ability to capture the attribute commonsense knowledge.

3.1 Problem Statement

Given a property p (e.g., color) and an object $o \in O$ (e.g., penguin), where O is a set of well-known objects, the task objective is rank K attribute values $C^p = \{c_1^p, ..., c_K^p\}$ (e.g., red, gray, ..., blue) according to their plausibility for the given object. c_i^p denotes the i-th candidate attribute of property p. The model's performance is evaluated by its ability to assign higher ranks to more plausible attribute values.

4 COMPASS

Our proposed method, i.e., COMPASS, estimates the plausibility of attribute values according to their influence on the semantics of a sentence. As shown in Figure 1b, COMPASS leverages the semantic shifts from a reference anchor sentence) describing the object to modified candidate sentences that include specific attribute values. This allows us to gauge how naturally each attribute fits with the object in a given context.

Property	Type	Templates
Color	Anch.	A photo of a [o].
	Cand.	A photo of a [c] [o].
Shape	Anch.	This is a [o].
	Cand.	This is a [o] with [c] shape.
Material	Anc.	This is an image of a [o].
	Cand.	This is an image of a [o] made of [c].

Table 1: Examples of templates we used for anchor (Anch.) and candidate (Cand.) sentences construction. The templates for CoDa are the same as Color.

4.1 Constructing Anchor and Candidate Sentences

Given the object o, property p, and candidate attribute set C^p, we first generate an anchor sentence s_0 and k candidate sentences $\{s_1, s_2, ..., s_k\}$ using predefined template pairs $(t_{\text{anch}}, t_{\text{cand}})$. The examples of templates are shown in Table 1, and a complete list of templates is in Appendix A. To construct an anchor sentence s_0, the template incorporates only the object without specifying the visual property:

$$s_0 = f(t_{\text{anch}}, o), \tag{1}$$

in which $f(t_{\text{anch}}, \cdot)$ represents making a sentence according to template t_{anch}. For example, when o = 'penguin' and t_{anch} = 'There is a [o].', the anchor sentence becomes s_0 = 'There is a penguin.' The anchor represents the average semantics of the object.

Similarly, to construct a candidate sentence s_i of attribute value c_i^p, we modify the template t_{cand} to include both the object and the attribute:

$$s_i = f(t_{\text{cand}}, o, c_i^p, p), 1 \leq i \leq k. \tag{2}$$

For instance, given t_{cand} = 'There is a [c] [o].', p = 'color', and c_i^p = 'red', the resulting candidate sentence would be s_i = 'There is a red penguin.' This allows us to systematically explore how dif-

ferent attributes affect the overall semantics when paired with the object.

4.2 Sentence Representation

We use a pre-trained LM θ to encode a sentence s_i into high-dimensional semantic representation r_i:

$$r_i = \theta(s_i). \tag{3}$$

By selecting different sentence representation methods according to model architecture, COMPASS can adapt to various backbones, ensuring its flexibility. We discuss different strategies for extracting sentence representations for both encoder-only models, such as CLIP (Radford et al., 2021) and RoBERTa (Liu et al., 2021c), and decoder-only models, such as Mistral (Jiang et al., 2023).

4.2.1 Encoder-Only Models

For models with an encoder-only architecture, we represent the entire sentence using the hidden state of the special '[CLS]' token, or the average hidden state of all tokens. Specifically, for a given sentence s of length l, which is prepended with the [CLS] token, the model θ outputs $l+1$ token hidden states $H = \{h_0, h_1, ..., h_{l+1}\}$:

$$H = \theta([CLS], s). \tag{4}$$

The final sentence representation r is derived from h_0 or $\text{avg}(H)$, depending on which one works better for a particular model.

4.2.2 Decoder-Only Models

In the case of decoder-only models, sentence representation can be more flexible. One common approach is to use the hidden state of the last token in the sentence s, which naturally encapsulates the context built up throughout the sequence:

$$H = \theta(s), \tag{5}$$

where $H = \{h_1, h_1, ..., h_l\}$ and r is derived from h_l, i.e., the last token hidden state.

Another alternative for is the PromptReps method (Zhuang et al., 2024), where the model is prompted to represent the input text using a single word. The hidden state of this generated word is then used as the sentence representation.

4.3 Ranking with Semantic Shifts

To rank an attribute value c_i^p concerning the object o and property p, we compute the semantic similarity between the anchor sentence representation r_0 and the candidate sentence representation r_i. Greater semantic similarity means smaller semantic deviation and higher plausibility. Therefore, the commonsense plausibility p_i of (o, p, c_i^p) is then defined as proportional to this similarity:

$$p_i(o, p, c_i^p) \propto sim(r_0, r_i), \tag{6}$$

where $sim(r_0, r_i)$ denotes a similarity function, such as cosine similarity or dot product. The larger the value of p_i, the higher c_i^p is ranked.

	color		shape		material	
	obj	cand	obj	cand	obj	cand
CoDa	521	11	-	-	-	-
ViComTe	2,877	12	706	12	1,423	18

Table 2: Number of objects (obj) and attribute candidates (cand) in CoDa and ViComTe.

5 Experimental Setup

5.1 Datasets

We evaluate the effectiveness of baseline methods and COMPASS using the following two datasets:

Color Dataset (CoDa) (Paik et al., 2021) is a human-annotated dataset that provides color distributions for commonly recognized objects. During crowd-sourcing, workers were presented with object names and asked to rate, on a scale from 1 to 5, the frequency of each object appearing in 11 predefined colors. The annotations are then normalized to form a probability distribution over the colors for each object.

Visual Commonsense Tests (ViComTe) (Zhang et al., 2022) is mined from the Visual Genome dataset(Krishna et al., 2017) and provides distributions for visual properties, i.e., color, shape, and material. The dataset was constructed by manually compiling a list of attributes for each property. Only objects in the Visual Genome that were matched with a specific attribute above a certain frequency threshold were included to minimize noise.

The number of objects and candidate attributes for each property is summarized in Table 2. They provide attribute distributions for common objects. Additionally, they include data for binary comparisons, i.e., (object, positive attribute, negative attribute). Objects in the dataset are further divided into three groups based on their attribute distribution: single, multi, and any. The 'single' group includes objects with one typical attribute (e.g.,

the color of snow). The 'multi' group consists of objects whose attributes are primarily distributed among the top four attribute values (e.g., the color of a penguin). The 'any' group represents objects with a broader attribute distribution (e.g., the color of a T-shirt).

5.2 Evaluation Metrics

We use **Spearman's rank correlation coefficient** ρ to evaluate the model's ranking performance. It measures the correlation between the predicted ranks of candidate attributes and their ground-truth ranks, focusing on the relative ordering rather than exact values. A ρ value of 1 indicates a perfect correlation, 0 indicates no correlation, and -1 indicates a perfect negative correlation.

For binary comparisons, we use **accuracy** as the evaluation metric, to assess whether the model can rank more common attributes higher.

5.3 Methods for Comparison

5.3.1 Baselines

Task-Specific Methods for Commonsense: We consider the following methods that are specifically designed to target commonsense.

COMET-Atomic-2020-Bart (abbreviated as **COMET**) (Bosselut et al., 2019) is a KB-driven LM that pre-trained on commonsense KB to generate tail entities a given head entity and relation. COMET generates a probability for each candidate being the tail entity, reflecting its alignment with the object's property. **ACCENT** (Ghazarian et al., 2023) is a KB-driven method that calculates the compatibility of tuples by comparing the maximum similarity between candidate attributes and the k tails generated by COMET. VERA-T5-XXL (abbreviated as **VERA**) (Liu et al., 2023) is a general-purpose model designed to estimate the commonsense plausibility of statements. It is trained on ~7M commonsense statements and formulates the task as a binary classification problem.

Plausibility Estimation with LMs: LMs are trained on extensive corpora, allowing them to internalize a large amount of commonsense knowledge. We consider the following several LMs.

We utilize **RoBERTa** (Liu et al., 2021c) a well-known LM has relatively small number of parameters. For LLM, we experiment with Mistral-7B-Instruct (Jiang et al., 2023) (abbreviated as **Mistral**), a powerful 7-billion-parameter model that demonstrates strong performance across a variety of instructions. We also test the performance of closed-source LLMs, i.e., **GPT-3.5** (OpenAI, 2022) (gpt-3.5-turbo-0125) and **GPT-4** (Achiam et al., 2023) (gpt-4-0125-preview).

5.3.2 COMPASS with Various Backbones

We evaluate the performance of COMPASS using both LMs and MLMs to determine its adaptability across different architectures.

CLIP-ViT-L/14 (abbreviated as **CLIP**) (Radford et al., 2021) is a multimodal representation model trained on extensive image-text pairs with contrastive learning. It encodes semantically similar images and text into closely aligned representations EVA-CLIP-8B-448. (abbreviated as **EVA-CLIP**) (Sun et al., 2023) builds upon CLIP by incorporating advanced representation learning techniques, optimization strategies, and enhancement methods, leading to improved performance.

We also adapt COMPASS to LMs. Specifically, we include RoBERTa-large (abbreviated as **RoBERTa + COMPASS**) (Liu et al., 2021c), a widely used LM. However, the sentence encoding capability of RoBERTa is limited, which is crucial for the effectiveness of COMPASS. Therefore, we incorporate two variants of RoBERTa, namely unsup-SimCSE and sup-SimCSE(abbreviated as **+ unsup-CL** and **+ sup-CL**) (Gao et al., 2021), which enhance sentence embeddings through contrastive learning in either an unsupervised or supervised manner. For LLMs, we test COMPASS on Mistral-7B-Instruct (abbreviated as **Mistral+ COMPASS**)(Jiang et al., 2023). Additionally, we evaluate a model based on Mistral that introduced contrastive learning to enhance sentence representation capability, i.e., E5-Mistral-7B-Instruct(abbreviated as **Mistral+ CL**) (Wang et al., 2023, 2022a), which aligns more closely with the requirements of COMPASS.

5.4 Implementation Details

For closed-source models, we obtain the generated results via their official APIs. For open-source models, all experiments are conducted on a single NVIDIA A800 80G GPU. Notably, no additional training is required for any of the experiments, as we focus on evaluating the models' pre-trained capabilities. We set $k = 10$ when evaluating ACCENT as in the original paper. For Both ACCENT and COMET, the object is treated as the head, candidate attribute is treated as the tail, and 'HasProperty' (a relation type in ConceptNet) is treated as the relation to adapt the task. We adapt VERA by

	Model	Parameters	CoDa	Color	Shape	Material
Baselines	ACCENT	440M	10.07	10.35	-2.10	16.99
	COMET	440M	22.91	26.98	40.44	25.72
	VERA	5B	58.93	45.08	30.31	33.51
	GPT-3.5	-	63.96	52.51	43.36	37.38
	GPT-4	-	65.53	54.42	42.91	36.00
LM	RoBERTa	355M	20.02	34.62	36.41	24.83
	+ COMPASS	355M	24.63	22.68	26.77	19.93
	+ unsup-CL & COMPASS	355M	32.67	32.00	42.18	31.12
	+ sup-CL & COMPASS	355M	44.59	38.92	42.92	33.55
LLM	Mistral	7B	46.64	38.63	30.46	36.34
	+ COMPASS	7B	53.14	39.07	39.75	31.88
	+ CL & COMPASS	7B	53.72	39.81	42.90	**38.83**
MLM	CLIP	124M	57.77	45.30	45.35	33.20
	EVA-CLIP	695M	**62.51**	**50.77**	**48.12**	38.68

Table 3: Test results on CoDa and ViComTe (Color, Shape, and Material). The best results among baselines and with COMPASS are underlined and bolded, respectively. All Spearman's ρ are scaled by a factor of 100. For each method, we experiment with several templates and report the results of the one template that achieves the overall best performance. '+ CL' denotes the models enhanced with contrastive learning. 'unsup' and 'sup' refer to unsupervised and supervised training. We also show the parameters of the text module, which is used for ranking.

Model	CoDa	Color	Shape	Material
GPT-3.5	94.05	92.25	90.08	89.60
GPT-4	94.63	**93.29**	89.24	88.76
Mistral	**95.97**	86.06	91.50	**91.27**
EVA-CLIP	95.39	**93.29**	**94.33**	90.79

Table 4: Binary comparison accuracy on CoDa and ViComTe. The best results are highlighted in bold. All results are scaled by a factor of 100. Both Mistral and EVA-CLIP use the COMPASS method.

using the candidate sentences as the input statements and rank candidates based on the output scores of each sentence, which can be used to estimate the plausibility of given statements. We utilize RoBERTa for commonsense estimation in a cloze task format, i.e., we construct sentences using predefined templates and compute the probability of each candidate filling in the blank. For Mistral and GPTs, we provide the object and property within the prompt and instruct the model to generate a ranked list of the attributes.

6 Results and Analysis

6.1 Overall Results

The results of baseline methods and our proposed COMPASS approach are summarized in Table 3. To provide a more intuitive understanding of the performance, we conducted additional binary comparison experiments on higher-performing models, with the results presented in Table 4. Several key observations can be made:

Current methods struggle with attribute commonsense. Both KB-driven approaches (ACCENT, COMET) and LLMs trained on large text corpora (VERA) face limitations. While KBs like ConceptNet include the 'HasProperty' relation, the limited coverage of triplets under this relation makes it inadequate for robust reasoning. Moreover, ACCENT fails to distinguish between similar candidates (e.g., red and green apples) due to its reliance on maximum similarity. VERA, despite fine-tuned on millions of sentences, struggles to generalize attribute commonsense, highlighting the insufficiency of relying solely on text data.

COMPASS consistently improves performance across various backbones. A comparison between RoBERTa and Mistral, with and without COMPASS, shows that COMPASS consistently enhances results. Since GPTs are closed-source models, we cannot adapt COMPASS to them. However, the improvement based on Mistral shows that our method is expected to improve their performance once the parameters can be accessed. Moreover,

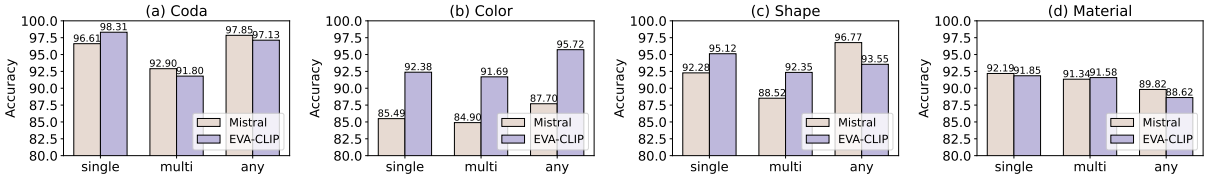


Figure 2: Binary classification accuracy of Mistral and EVA-CLIP with COMPASS on different datasets. The result of each group is shown individually.

while previous works suggested that CLIP, not being trained for token-prediction, was unsuitable for knowledge access without fine-tuning (Paik et al., 2021; Zhang et al., 2022), our approach effectively exploits the knowledge in the CLIP without requiring additional training, demonstrating the model's latent capacity to support commonsense reasoning.

Sentence representation quality is crucial for COMPASS. Since COMPASS relies on measuring semantic similarity, the quality of sentence representations and the model's capacity to capture semantic differences are essential. For models with weaker representation capabilities, such as RoBERTa and Mistral, the direct application of COMPASS may yield inconsistent results. However, incorporating contrastive learning to enhance the sentence representation ability can substantially improve their effectiveness in our method. Notably, even unsupervised training can yield strong results with COMPASS, underscoring that the Improvement hinges on representation quality, rather than relying on additional data.

Multimodal models outperform unimodal models in capturing attribute commonsense. With the integration of COMPASS, EVA-CLIP delivers results comparable to, or even surpassing, LLMs, including advanced closed-source models like GPT-3.5 and GPT-4, despite having significantly fewer parameters on the text side. Visual data inherently contains rich commonsense knowledge, particularly regarding object attributes. Consequently, even though models like CLIP and EVA-CLIP are not specifically trained to model sentence likelihood, multimodal supervision allows them to internalize substantial commonsense knowledge. This highlights the efficiency of leveraging visual information to enhance commonsense learning.

Verbalization may not always be optimal for LLMs. Although LLMs are inherently well-suited for generative tasks, relying on verbalization approaches may not always be more effective than representation-based methods. As our experiments demonstrate, designing methods that leverage a model's internal representations can lead to more accurate results. This suggests that carefully exploiting the model's latent knowledge through representation-based techniques, rather than solely depending on the generated text, may yield superior outcomes in commonsense estimation tasks.

6.2 Further Analyses

We analyze the performance of Mistral+COMPASS (hereafter referred to as Mistral) and EVA-CLIP across different settings and data types to further explore the differences when COMPASS is applied to various model architectures.

6.2.1 Visual Information is More Helpful for Single Group

To compare performance across different types of data, we do binary comparisons among different groups. The results are shown in Figure 3. We observe that MLMs outperform LMs mainly in the single group, indicating that visual information is particularly beneficial for modeling typical attributes. This may be because the more typical attributes are more commonly accepted by people as a consensus, and therefore are less likely to be mentioned in communications, making it harder for LMs to capture this type of knowledge. In contrast, these attributes are clearly depicted in images, allowing MLMs to better learn and represent them by incorporating visual information.

6.2.2 Importance of Template Format

An intuitive approach for estimating semantic shifts is to directly compare word collocations. For instance, when ranking the color of penguins, one might simply compare 'penguin' and 'black penguin' without constructing full sentences. The results of both methods are shown in Figure 3 (a). The results indicate that constructing complete sentences is crucial for COMPASS to perform well, regardless of whether the model is unimodal or multimodal. This is because sentence-level input more closely resembles the data format seen during

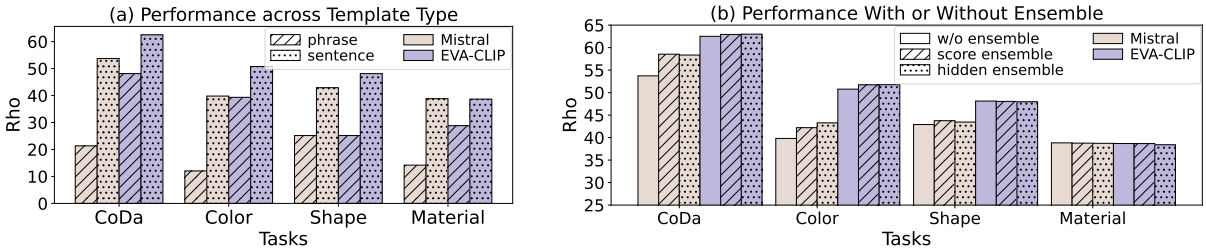


Figure 3: COMPASS performance on E5-Mistral and EVA-CLIP with different setting.

Sort the candidate shapes based on how frequently cakes are observed in each shape.

Ground-truth: **round (68%), square (23%), heart (4%), triangle (3%), rectangle (2%)**
GPT-3.5 : **round, rectangle, square, star, heart, cross, polygon, triangle, oval, semicircle, rhombus, octagon**
GPT-4 : **rectangle, round, square, heart, star, oval, polygon, triangle, rhombus, semicircle, octagon, cross**
Mistral : **oval, round, semicircle, square, star, cross, heart, triangle, polygon, octagon, rhombus, rectangle**
Mistral* : **round, rectangle, square, oval, heart, star, triangle, semicircle, polygon, cross, rhombus, octagon**
EVA-CLIP* : **round, square, triangle, rectangle, heart, oval, star, cross, octagon, semicircle, polygon, rhombus**

Figure 4: The ground-truth distribution of cake shapes and the ranking results of different models. The darker the shape, the higher it should be ranked. The '*' in the upper right represents the model with COMPASS method.

model pre-training. Additionally, complete semantics play a particularly important role for LMs, as they typically encounter longer texts during pre-training (e.g., documents), compared to the shorter inputs, like captions, often seen by MLMs.

6.2.3 Templates Ensemble is more Effective for Language Models

We analyzed the impact of using multiple templates ensemble. Specifically, the score ensemble combines the scores from different templates to generate the final prediction. The hidden ensemble fuses the sentence representations before calculating the final score. As shown in Figure 3 (b), both ensemble strategies significantly improve performance for Mistral, particularly in the CoDa and Color tasks, but offer limited benefit for EVA-CLIP. This may be because LM relies heavily on textual input and benefits from the diverse linguistic contexts provided by multiple templates. In contrast, the sentence structure in MLM pretraining data is relatively simpler, resulting in more consistent performance even with a single template.

6.3 Case Study

We conducted a case study to qualitatively analyze the performance of different models, using the shape of cakes as an example (see Figure 4). For the most common shape, round, we observed that the LM does not always rank it in the top position. Additionally, despite triangular cakes being quite common, the LM consistently ranks triangles lower than expected. In contrast, the MLM EVA-CLIP produces more reasonable rankings, highlighting the presence of reporting bias in text data and the value of incorporating visual information. When comparing Mistral's direct prediction rankings with those using COMPASS, we find that COMPASS allows for more accurate extraction of the model's knowledge, further demonstrating the effectiveness of our method.

7 Conclusion

In this work, we propose COMPASS, a novel method designed to estimate attribute commonsense plausibility through semantic shifts in sentence representations. By leveraging the semantic shifts between anchor and candidate sentences, COMPASS allows for a more nuanced understanding of how plausible an attribute is for a given object. Our extensive experiments demonstrated that COMPASS consistently enhances performance across various model architectures, including LMs and MLMs, without fine-tuning requirements. We also find that Multimodal models like EVA-CLIP, which integrate visual information, show higher efficiency in capturing attribute commonsense than LMs. Additionally, the importance of sentence representation quality was highlighted, with contrastive learning significantly improving the effectiveness of models with weaker sentence encoding capabilities. Through COMPASS, we show that integrating visual information and leveraging semantic shifts is a promising direction for commonsense plausibility estimation.

8 Limitations and Ethical Considerations

Our method relies on predefined templates to construct sentences for objects and candidate attributes. Automating template generation could be an important avenue for future improvement. As previous methods apply to specific types of commonsense, such as event commonsense knowledge, our method specifically focuses on attribute commonsense. Future work should explore integrating various methods to cover a broader range of commonsense knowledge.

As our method relies on LLMs and MLMs, it inherits potential biases present in the training data. These biases, whether related to societal stereotypes or uneven distribution of information across certain attributes, could affect the model's judgment in ranking attribute plausibility. Consequently, our method may inadvertently perpetuate or amplify these biases, especially in scenarios where the model's understanding of an attribute is skewed by biased representations in the data. Addressing these biases is an important avenue for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Annual Meeting of the Association for Computational Linguistics*.

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023. Say what you mean! large language models speak too positively about negative commonsense knowledge. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Byeongmin Choi, Yong-Sook Lee, Yeunwoong Kyung, and Eunchan Kim. 2022. Albert with knowledge graph encoder utilizing semantic similarity for commonsense question answering. *ArXiv*, abs/2211.07065.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. MORE: Multi-mOdal REtrieval augmented generative commonsense reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1178–1192, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pretrained models. 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)*, pages 1173–1178.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Steven Y Feng, Kevin Lu, Zhuofu Tao, Malihe Alikhani, Teruko Mitamura, Eduard Hovy, and Varun Gangal. 2022. Retrieve, caption, generate: Visual grounding for enhancing commonsense in text generation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10618–10626.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Sarik Ghazarian, Yijia Shao, Rujun Han, A. G. Galstyan, and Nanyun Peng. 2023. Accent: An automatic event commonsense evaluation metric for open-domain dialogue systems. In *Annual Meeting of the Association for Computational Linguistics*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Conference on Automated Knowledge Base Construction*.

Jerry R Hobbs. 1990. Granularity. In *Readings in qualitative reasoning about physical systems*, pages 542–545. Elsevier.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4*, AS '02, pages 45–51.

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*, pages 1264–1287.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021a. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6418–6425.

Yixian Liu, Liwen Zhang, Wenjuan Han, Yue Zhang, and Kewei Tu. 2021b. Constrained text generation with global guidance–case study on commongen. *arXiv preprint arXiv:2103.07170*.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021c. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Gary F. Marcus. 2020. The next decade in ai: Four steps towards robust artificial intelligence. *ArXiv*, abs/2002.06177.

Ira Noveck and Dan Sperber. 2004. *Experimental pragmatics*. Springer.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The world of an octopus: How reporting bias influences a language model's perception of color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Roger C Schank. 1983. *Dynamic memory: A theory of reminding and learning in computers and people*. cambridge university press.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

Yufei Tian, Felix Zhang, and Nanyun Peng. 2023. Harnessing black-box control to boost commonsense in lm's generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5417–5432.

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

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. 2021. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

PeiFeng Wang, Jonathan Zamora, Junfeng Liu, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022b. Contextualized scene imagination for generative commonsense reasoning. In *International Conference on Learning Representations*.

T Winograd. 1986. Understanding computers and cognition: A new foundation for design.

W. Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Conference on Empirical Methods in Natural Language Processing*.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022. Visual commonsense in pretrained unimodal and multimodal models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5321–5335.

Yi Zhang, Lei Li, Yunfang Wu, Qi Su, and Xu Sun. 2021. Alleviating the knowledge-language inconsistency: A study for deep commonsense knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:594–604.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2024. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2024. Promptreps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. *arXiv preprint arXiv:2404.18424*.

A Templates for Sentence Construction

The templates we used to construct anchor sentences and candidate sentences of different property are shown in Table A.

Property	Templates for anchor	Templates for candidate
Color	A photo of a [o].	A photo of a [c] [o].
	A picture of a [o].	A picture of a [c] [o].
	An image of a [o].	An image of a [c] [o].
	An image of a [o].	An image of a [o] which is [c].
	There is an image of a [o].	There is an image of a [c] [o].
	There is a photo of a [o].	There is a photo of a [c] [o].
	There is a picture of a [o].	There is a picture of a [c] [o].
	There is an image of a [o].	There is an image of a [o] which is [c].
	There is a photo of a [o].	There is a photo of a [o] which is [c].
	It is an image of a [o].	It is an image of a [o] which is [c].
	It is a photo of a [o].	It is a photo of a [o] which is [c].
	There is a [o].	There is a [o] in [c].
	There is a [o].	There is a [o] which is [c].
	Everyone knows [o].	Everyone knows that [o] is [c].
	Everyone knows [o].	Everyone knows that [o] is [c].
Shape	This is a [o].	This is a [o] with [c] shape.
	There is a [o].	There is a [c] [o].
	There is a [o].	There is a [o] which shape is [c].
	It is an image of a [o].	It is an image of a [o] which shape is [c].
	There is an image of a [o].	It is an image of a [o] which shape is [c].
	There is an image of a [o].	There is an image of a [c] [o].
	There is a picture of a [o].	There is a picture of a [c] [o].
	There is a picture of a [o].	There is an picture of a [o] which shape is [c].
	There is a picture of a [o].	There is an picture of a [c] [o].
	This is a picture of a [o].	This is a picture of a [o] has [c] shape.
	A picture of a [o].	A picture of a [o] has [c] shape.
	An image of a [o].	An image of a [c] [o].
	A photo of a [o].	A photo of a [c] [o].
	A picture of a [o].	A picture of a [c] [o].
	[o] is of shape .	[o] is of shape [c].
	The shape of [o].	The shape of [o] can be [c].
	The shape of the [o].	The shape of the [o] is [c].
Material	This is an image of a [o].	This is an image of a [o] made of [c].
	This is an image of a [o].	This is an image of a [o] which made from [c].
	This is an image of a [o].	This is an image of a [o] which made of [c].
	This is a photo of a [o].	This is a photo of a [o] made of [c].
	This is a picture of a [o].	This is a picture of a [o] made of [c].
	This is a picture of a [o].	This is a picture of a [o] which made of [c].
	It is a picture of a [o].	It is a picture of a [o] made of [c].
	A picture of a [o].	A picture of a [o] which made from [c].
	A picture of a [o].	A picture of a [o] which made of [c].
	A picture of a [o].	A picture of a [c] [o].
	There is an image of a [o].	There is an image of a [c] [o].
	There is a photo of a [o].	There is an photo of a [c] [o].
	There is a picture of a [o].	There is an picture of a [c] [o].
	An image of a [o].	An image of a [c] [o].
	A photo of a [o].	A photo of a [c] [o].
	A picture of a [o].	A picture of a [c] [o].

Table 5: Templates we used for constructing anchor sentences and candidate sentences. The templates for CoDa are the same as Color.