Estimating Commonsense Plausibility through Semantic Shifts

Anonymous ACL submission

Abstract

Commonsense plausibility estimation is critical for evaluating language models (LMs), yet existing generative approaches–reliant on likelihoods or verbalized judgments–struggle with fine-grained discrimination. In this paper, we propose ComPaSS, a novel discriminative framework that quantifies commonsense plausibility by measuring semantic shifts when augmenting sentences with commonsense-related information. Plausible augmentations induce minimal shifts in semantics, while implausible ones result in substantial deviations. Evaluations on two types of fine-grained commonsense plausibility estimation tasks across different backbones, including LLMs and vision-language models (VLMs), show that ComPaSS consistently outperforms baselines. It demonstrates the advantage of discriminative approaches over generative methods in fine-grained commonsense plausibility evaluation. Experiments also show that (1) VLMs yield superior performance to LMs, when integrated with ComPaSS, on vision-grounded commonsense tasks. (2) contrastive pre-training sharpens backbone models' ability to capture semantic nuances, thereby further enhancing ComPaSS.

1 Introduction

Commonsense knowledge–the shared understanding of everyday phenomena and human experiences (Schank, 1983; Winograd, 1986; Hobbs, 1990)–is foundational to natural language understanding and generation. Despite the remarkable progress in large language models' (LLMs) text generation capabilities, ensuring commonsense plausibility in their outputs remains an unresolved challenge (Marcus, 2020; Elazar et al., 2021; Mahowald et al., 2024; Chen et al., 2023). This challenge arises not only from the inherent difficulty of acquiring and applying commonsense knowledge but also from the absence of reliable frameworks for evaluating textual plausibility. Effective evaluation of commonsense plausibility addresses this gap twofold: it identifies commonsense violations while offering quantifiable metrics to guide the development of techniques that augment LLM outputs.

In this work, we focus on developing generalizable methods for commonsense plausibility estimation (CSPE) that can be applied across diverse domains and tasks. This leads us to investigate zero-shot and few-shot approaches based on pre-trained LMs, which leverage their inherent knowledge without requiring additional training data or domain-specific fine-tuning.

Previous studies on zero or few-shot CSPE primarily adopt a generative perspective and can be categorized into two main approaches, likelihood estimation and verbalized judgments. The likelihood-based methods (Trinh and Le, 2018; Tamborrino et al., 2020; Holtzman et al., 2021) utilize token prediction probabilities from language models as an indicator, with the assumption that sentences consistent with commonsense knowledge tend to have a higher likelihood for their component tokens. The verbalization-based methods (Brown et al., 2020; Krause and Stolzenburg, 2024) ask pre-trained LMs to answer the plausibility of a sentence through natural language. The models can generate the answer based on knowledge stored in their parameters.

However, approaches based on the generative perspective could be suboptimal for CSPE, since it is essentially a discriminative task. In this paper, we adopt a discriminative perspective for CSPE. In communication, commonsense knowledge is often assumed and left unstated, yet such omissions rarely hinder mutual understanding (Clark, 1996; Noveck and Sperber, 2004). Inspired by this, we propose ComPaSS, a method that measures **Com**monsense **Pla**usibility through **S**emantic **S**hifts introduced when augmenting sentences with commonsense-related information. Plausible additions yield minimal semantic shifts, whereas implausible ones result in substantial deviations. For instance, adding ‘black’ to ‘There is a penguin’ results in a minor semantic shift, aligning with the penguins’ natural coloration. By contrast, introducing ‘green’ creates a substantial shift, highlighting the implausibility of such an atypical attribute. To quantify semantic shifts, ComPaSS computes the similarity between embeddings of the original sentence (without explicit commonsense references) and its modified counterpart augmented with commonsense-related information.

Two aspects of semantic representations could influence the capability of ComPaSS in CSPE: the inclusion of commonsense knowledge and the discrimination of semantic nuances. These correspond to two key aspects of models used for obtaining sentence embeddings: 1) Modality. Language Models (LMs) often suffer from *reporting bias* (Gordon and Durme, 2013), which involves systematic distortions due to omitted commonsense details (e.g., ‘penguins are black’ is rarely stated) and statistical biases from fixed linguistic patterns (e.g., ‘black sheep’). In contrast, vision-language models (VLMs) incorporate visual information, thus mitigating reporting bias, especially for visually-grounded commonsense knowledge (e.g., object colors or spatial relations) (Paik et al., 2021; Zhang et al., 2022). 2) Contrastive learning. By training a model to distinguish between semantically similar and dissimilar instances, it enhances the model’s discriminative power. Representations from contrastively trained models exhibit sharper separability, which directly impacts the precision of semantic shift measurements. Given these considerations, we study how ComPaSS performs based on various backbones of both LMs and VLMs, with and without contrastive learning.

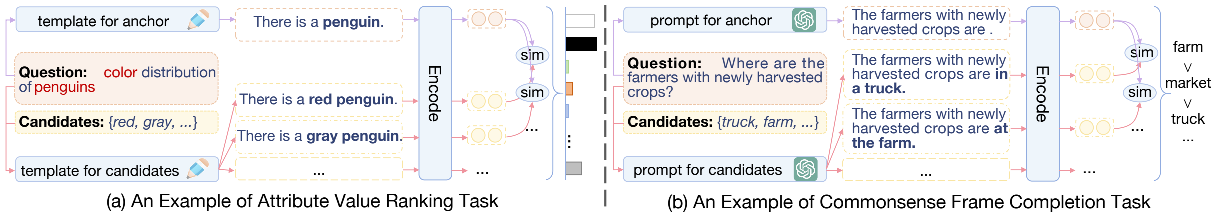
We evaluate ComPaSS against baselines on two fine-grained CSPE tasks that require ranking candidate answers by plausibility rather than binary classification. These tasks prioritize nuanced plausibility judgments, where answers may hold varying degrees of validity. The first task, attribute value ranking, involves ranking candidate attribute values (e.g., color, shape, material) for objects using structured triplets (e.g., determining that "black" is more plausible than "green" for penguin-color), evaluated on datasets like CoDa (Paik et al., 2021) and ViComTe (Zhang et al., 2022). The second task, commonsense frame completion (Cheng et al., 2024), challenges models to rank plausible completions for open-ended prompts (e.g., selecting ‘farm’ over ‘truck’ for ‘Where are farmers with newly harvested crops?’), testing alignment with human preferences and broader commonsense reasoning. Together, these tasks assess ComPaSS across input formats (structured triplets vs. free-form text) and knowledge types (object-specific attributes vs. contextual, inferential commonsense).

Our experiments reveal three critical insights. First, as a discriminative approach, ComPaSS consistently outperforms prior generative methods in fine-grained plausibility estimation, achieving superior results across diverse model backbones. This highlights the advantage of discriminative methods in capturing subtle plausibility distinctions. Second, utilizing ComPaSS, VLMs significantly outperform LMs for vision-grounded commonsense (e.g., object colors or shapes), demonstrating that visual information enhances representations and benefits CSPE. Third, models with contrastive pretraining yield significantly better results than those without, emphasizing the importance of representations that capture semantic nuances in plausibility measurement through ComPaSS.

2 Related Work

2.1 CSPE Based on Internal Knowledge

The sentence probability and perplexity computed by LMs can serve as indicators of commonsense plausibility, even in zero-shot settings (Trinh and Le, 2018; Davison et al., 2019; Liu et al., 2021a). For LLMs with instruction-following capability, they can be directly prompted to judge whether a given input is consistent with commonsense or not (Zhao et al., 2024). Beyond directly judging plausibility, 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 LLMs, i.e., the reasoning chains justifying ‘reasonable’ or ‘unreasonable’ conclusions, and selecting the final prediction based on the highest-scoring path. VERA (Liu et al., 2023) adopts a discriminative approach, training a classification head to make predictions based on model representations, which fine-tunes LLMs on~7 million commonsense statements. In contrast, our approach also leverages internal knowledge from a discriminative perspective but does not require additional training.


template for anchor
There is a penguin.
Question: color distribution of penguins
There is a red penguin.
Candidates: {red, gray, ...}
There is a gray penguin.
template for candidates
...
Encode
sim
sim
...
prompt for anchor
The farmers with newly harvested crops are .
Question: Where are the farmers with newly harvested crops?
The farmers with newly harvested crops are in a truck.
Candidates: {truck, farm, ...}
The farmers with newly harvested crops are at the farm.
prompt for candidates
...
Encode
sim
sim
farm
>
market
>
truck
...

(a) An Example of Attribute Value Ranking Task

(b) An Example of Commonsense Frame Completion Task

Figure 1: How ComPaSS works on different tasks.

2.2 CSPE Based on External Knowledge

Language models (LMs) may have insufficient or inaccurate knowledge, which led to some methods to incorporate external knowledge to better estimate commonsense plausibility. A typical approach is to augment the model's knowledge by retrieving relevant sentences from external sources (Zhang et al., 2021; Yu et al., 2022). Commonsense knowledge bases (KBs) (Speer et al., 2016; Sap et al., 2019; Hwang et al., 2020) store extensive commonsense knowledge, enabling the extraction of relevant subgraphs to evaluate sentence consistency with commonsense (Choi et al., 2022). To alleviate the coverage limitations of the KBs while leveraging the extensive knowledge encoded in LMs, COMET (Bosselut et al., 2019) introduced a dynamic KB by pre-training LM on existing commonsense KBs. Methods that utilize this dynamic KB (Ghazarian et al., 2023; Tian et al., 2023) demonstrate improved generalization across various commonsense reasoning tasks.

3 Task Definition

Formally, given an input instance $x_i = (c; a_i^c)$ consisting of a context c and a candidate information $a_i^c \in A$, where $A^c = \{a_1^c, a_2^c, ..., a_K^c\}$ denotes the context-dependent candidate set with size K, the task is to predict a plausibility score set $\mathcal{P}^c = \{p_1^c, p_2^c, ..., p_K^c\}$ for all candidates, where each $p_i^c \in \mathbb{R}$ quantifies the plausibility of augmenting c with a_i^c. The ground-truth scores are denoted as $\mathcal{G}^c = \{g_1^c, g_2^c, ..., g_K^c\}$, where g_i^c indicates the true score of a_i^c. Performance is measured by the correlation between $\mathcal{P}^c$ and $\mathcal{G}^c$.

The input can take two specific forms: for *attribute value ranking* task, the input is a structured triplet $x_i = (o, \text{has property } p; a_i^c)$. The context $c = (o, \text{has property } p)$, where o is a common object and p is a property. The candidate a_i^c represents the i-th attribute value for the specified property. For the *commonsense frame completion* task, the context $c = q$ is a free-form question, the input is a question-answer pair $x_i = (q; a_i^c)$, where a_i^c is the i-th plausible answer to this question.

4 ComPaSS

Our method, ComPaSS, is a zero-shot approach for estimating commonsense plausibility. We demonstrate in Figure 1 how this method works on different tasks. For each input, we first construct an anchor sentence (omitting the commonsense-related detail) and a candidate sentence (augmenting that detail). We then encode both sentences individually to obtain their semantic representations. Next, we calculate their semantic similarity, where the degree of semantic shift—inversely proportional to similarity—quantifies plausibility.

4.1 Constructing Sentences

For each input context c and the candidate to be evaluated a_i^c, we construct two types of sentences: an anchor sentence s_{anchor} that contains only the base context c while omitting target details, and a candidate sentence s_{candi} that further incorporates commonsense-related information a_i^c. The construction process varies based on input type but follows a unified framework:

$$s_{\text{anchor}} = f_{\text{anchor}}(c, z_{\text{anchor}}), \quad (1)$$

$$s_{\text{candi}} = f_{\text{candi}}(c, a_i^c, z_{\text{candi}}), \quad (2)$$

where $f(\cdot) \in \{f_{\text{anchor}}(\cdot), f_{\text{candi}}(\cdot)\}$ denotes the construction function, and $z \in \{z_{\text{anchor}}, z_{\text{candi}}\}$ denotes task-specific templates or prompts.

As illustrated in Figure 1, the framework is instantiated differently based on the input format: For *triplet inputs*, we employ template-based construction, where z represents a pre-defined template (see Appendix A) and $f(\cdot)$ represents applying this template to generate a sentence. For *question-answer pairs*, we query GPT-4 (Achiam et al., 2023) for sentence transformation, where z denotes the prompt (see Appendix B) and $f(\cdot)$ represents query GPT-4 using the specified prompt. Since questions cannot be directly converted into coherent statements, we use a blank space as a placeholder when constructing anchor sentences.

4.2 Representing Sentences

Given anchor and candidate sentences, we encode them into dense semantic representations using a pre-trained model θ, which can be either a LM or a VLM. For each sentence $s \in \{s_{\text{anchor}}, s_{\text{candi}}\}$, the model first processes the sentence along with special tokens (e.g., [CLS], [EOS], or others depending on the model architecture) and then outputs token hidden states:

$$H = \theta(s) = \{h_0, h_1, ..., h_l\}, \tag{3}$$

where l denotes the sequence length, including the special tokens. The final sentence representation $r \in \{r_{\text{anchor}}, r_{\text{candi}}\}$ is derived through architecture-specific strategies.

For encoder models, we use the hidden state of the designated semantic aggregation token as sentence representation. Some models (e.g., RoBERTa (Liu et al., 2021b)) use the initial '[CLS]' token for sentence representation ($r = h_0$), while others (e.g., CLIP (Radford et al., 2021)) utilize the final '[EOS]' token embedding ($r = h_l$).

For decoder models, we use the hidden state of the last token as sentence representation $r = h_l$, which naturally encapsulates the accumulated context. Alternatively, PromptReps (Zhuang et al., 2024) prompts the model to generate a new representative token at position $l + 1$, using its hidden state as the sentence representation ($r = h_{l+1}$). We apply this strategy to models that are not enhanced by contrastive learning.

This architecture-aware representation strategy ensures ComPaSS's flexibility across different model backbones while maintaining optimal performance for each specific architecture.

4.3 Ranking with Semantic Shifts

We rank the candidate option a_i^c by measuring how naturally it integrates into the context, quantified through semantic similarity between the anchor sentence representation r_{anchor} and the candidate sentence representation r_{candi}. The underlying principle is that the more plausible the information, the smaller the semantic shifts it induces when added to the context, leading to higher semantic similarity. Formally, we define the commonsense plausibility score p_i^c for each candidate a_i^c as:

$$p_i^c \propto \text{sim}(r_{\text{anchor}}, r_{\text{candi}}), \tag{4}$$

where $\text{sim}(\cdot)$ denotes a similarity function (e.g., cosine similarity or dot product). Candidates are then ranked by their plausibility scores descendingly, with higher-ranked candidates representing more commonsense-consistent answers.

4.4 Discussion of Applicable LMs

This paragraph discusses the differences in applicable LMs between ComPaSS and generative methods based on likelihoods and verbalization. ComPaSS can utilize both encoder and decoder style models as long as they can yield reasonable sentence representations. Likelihood-based approaches can also leverage these two types of LMs. Candidate likelihoods can be estimated based on masked/next token prediction for encoders and decoders respectively. In contrast, verbalization-based approaches require LLMs–decoder-only LMs–to answer the plausibility estimation questions. This indicates the broader applicability of ComPaSS.

5 Experimental Setup

5.1 Datasets

We evaluate methods through two types of fine-grained commonsense plausibility estimation (CSPE) tasks, where candidates should be ranked based on commonsense plausibility. These tasks are carefully chosen to comprehensively evaluate methods across varying input formats (from structured triplets to free-form text) and commonsense knowledge levels (from specific attribute knowledge to general commonsense knowledge).

5.1.1 Structured Attribute Knowledge

Color Dataset (CoDa) [1] (Paik et al., 2021) is a human-annotated dataset used for attribute value ranking, which provides color distributions for commonly recognized objects. It contains 521 objects, each with 11 candidate color attributes.

Visual Commonsense Tests (ViComTe) [2] (Zhang et al., 2022) is another dataset used for attribute value ranking, which derived from the multimodal dataset, i.e., Visual Genome (Krishna et al., 2017). It offers attribute value distributions across a broader set of properties, including color, shape, and material. It contains 2,877 objects with 12 candidate color attributes, 706 objects with 12 candidate shape attributes, and 1,423 objects with 18 candidate material attributes.

[1] https://github.com/nala-cub/coda

[2] https://github.com/ChenyuHeidiZhang/VL-commonsense

5.1.2 Free-form General Knowledge

Commonsense Frame Completion (CFC) [3] (Cheng et al., 2024) is a dataset designed to evaluate implicit commonsense reasoning, which consists of questions accompanied by multiple plausible answers with human-annotated preference scores. It requires models to make probabilistic judgments about answer plausibility. The evaluation protocol employs a probabilistic framework that measures how well a model's predicted answer distribution aligns with human preferences. As the test set is not public, we use the validation set containing 55 questions for zero-shot evaluation.

5.2 Evaluation Metrics

Spearman's rank correlation coefficient ρ: We choose this as the primary metric following CoDa and ViComTe. It measures the correlation between the predicted ranks of candidates and their ground-truth ranks, focusing on the relative ordering rather than exact values. This emphasis on relative ordering aligns with the nature of commonsense plausibility assessment, where the exact probability values are less important than correctly identifying more plausible options over less plausible ones. A ρ value of 1 indicates perfect correlation, 0 indicates no correlation, and -1 indicates perfect negative correlation.

Accuracy: CoDa and ViComTe include binary comparison tasks where each object is paired with two attribute values, with one being more plausible than the other. The model need to rank the more plausible value higher. We use accuracy as the evaluation metric, which measures the proportion of correct rankings. This metric is particularly suitable for cross-attribute comparisons as it is unaffected by variations in the number of candidates, unlike the Spearman's rank correlation coefficient.

5.3 Methods for Comparison

5.3.1 ComPaSS with Various Backbones

We evaluate ComPaSS across diverse model architectures to assess its adaptability:

For LMs, we consider various open-source models, including RoBERTa-Large (Liu et al., 2021b) (RoBERTa), a widely-used encoder-only LM, along with two decoder-only LLMs, Mistral-7B-Instruct (Jiang et al., 2023) (Mistral) and Qwen2-7B-instruct (qwe, 2024) (Qwen2), both demonstrating strong instruction-following capabilities. We also evaluate their **contrastive learning-enhanced** variants, i.e., sup-SimCSE-RoBERTa-Large (Gao et al., 2021) (RoBERTa$_{\text{w/ CL}}$), E5-Mistral-7B-Instruct (Wang et al., 2023, 2022) (Mistral$_{\text{w/ CL}}$) and gte-Qwen2-7B-instruct (Li et al., 2023) (Qwen2$_{\text{w/ CL}}$).

For VLMs, we test CLIP-ViT-L/14 (Radford et al., 2021) (CLIP), a multimodal representation model trained on image-text pairs using **contrastive learning**, which aligns semantically similar images and text into closely matching representations. We also consider its advanced variant EVA-CLIP-8B (Sun et al., 2023) (EVA-CLIP) with improved performance.

5.3.2 Baselines

We compare against two categories of baseline methods:

Commonsense models (CSMs): These models are specifically designed for modeling commonsense knowledge: COMET-Atomic-2020-Bart (Bosselut et al., 2019) (COME-Atomic) is a commonsense LM pre-trained on commonsense KBs. COMET is suitable for processing triple input, which can generate a probability score for each candidate. ACCENT (Ghazarian et al., 2023) assesses the commonsense plausibility of a sentence by first extracting structured tuples and then scoring them based on their compatibility with a commonsense KB. VERA-T5-XXL (Liu et al., 2023) (VERA-T5) is trained on ~7M commonsense statements and can directly estimate the commonsense plausibility of statements.

Language models (LMs): We evaluate all open-source LMs used as the backbone of ComPaSS with both likelihood-based and verbalization-based approaches. For the *likelihood-based* method, the plausibility of a sentence is determined by the probability of predicting each token in the sentence sequentially, normalized by sentence length. A higher probability indicates greater plausibility. In the case of *verbalization-based* method, pre-trained language models are prompted in natural language (see Appendix C) to rank all candidate responses based on their plausibility. We also test closed-source LLMs including gpt-3.5-turbo-0125 (OpenAI, 2022) (GPT-3.5) and gpt-4-0125-preview (Achiam et al., 2023) (GPT-4), the latter introduces multi-modality and has superior capabilities.

[3] https://github.com/qxc101/PROBEVAL_CFC/

	Model (#Inference Parameters)	CoDa	Color	Shape	Material	CFC
	Baselines					
CSM	ACCENT (440M)	10.07	10.35	-2.10	16.99	35.04
	COMET-Atomic (440M)	22.91	26.98	40.44	25.72	-
	VERA-T5 (5B)	58.93	45.08	30.31	33.51	45.81
LM	RoBERTa+likelihood (355M)	24.37	33.63	36.12	24.23	42.46
	RoBERTa$_{\text{w/ CL}}$+likelihood (355M)	23.36	31.51	26.69	22.23	38.03
	Mistral+verbal. (7B)	46.64	38.63	30.46	36.34	32.06
	Mistral+likelihood (7B)	51.30	34.31	26.70	37.03	47.98
	Mistral$_{\text{w/ CL}}$+likelihood (7B)	25.70	4.72	18.81	5.96	35.46
	Qwen2+verbal. (7B)	57.40	41.59	38.30	36.76	29.32
	Qwen2+likelihood (7B)	50.25	40.99	32.52	37.13	45.10
	Qwen2$_{\text{w/ CL}}$+likelihood (7B)	49.65	41.75	32.80	37.30	43.00
	ComPASS					
LM	RoBERTa$_{\text{w/ CL}}$ (355M)	44.59	38.92	42.92	33.55	44.46
	Mistral$_{\text{w/ CL}}$ (7B)	58.54	42.20	43.75	**38.77**	**49.01**
	Qwen2$_{\text{w/ CL}}$ (7B)	$\underline{59.16}$	44.61	$\underline{47.51}$	38.49	46.41
VLM	CLIP (124M)	58.10	$\underline{45.55}$	45.82	33.56	35.13
	EVA-CLIP (695M)	**62.87**	**51.73**	**48.05**	$\underline{38.67}$	41.46

Table 1: Spearman's rank correlation coefficient ρ between the predicted ranks of candidates and their ground-truth on CoDa, ViComTe (Color, Shape, and Material), and CFC, shown in percentage. The **best** and $\underline{\text{second best}}$ results are highlighted in bold and underlined, respectively. '+verbal.' indicates using the verbalization-based method.

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, all experiments are carried out in a zero-shot or in-context few-shot setting. For ACCENT, we set the beam number to 10 as specified in the original paper. When testing on the CFC dataset using the verbalization method, we sample the model 100 times for each question with a temperature of 0.7.

Method	CoDa	Color	Shape	Material
likelihood	24.37	33.63	36.12	24.23
ComPaSS	24.63	22.68	26.77	19.93
w/ unsup-CL	32.67	32.00	42.18	31.12
w/ sup-CL	44.59	38.92	42.92	33.55

Table 2: Performance of different Roberta variants. By default we use the vanilla RoBERTa. 'w/ unsup-CL' and 'w/ sup-CL' denote RoBERTa pre-trained with unsupervised and supervised contrastive learning, respectively.

6 Results and Analysis

6.1 Overall Results

The overall experimental results comparing baseline methods with our approach are presented in Table 1, which reveals several key findings:

ComPaSS achieves the best performance compared to baselines. Further comparison between RoBERTa, Mistral, and Qwen2, with and without ComPaSS, shows a consistent improvement when ComPaSS is applied. This validates our method's architecture-agnostic effectiveness. Notably, even VERA, which was specifically fine-tuned for CSPE, achieves only comparable performance to ComPaSS-enhanced models. Comparing the performance of different methods on LMs in the baseline, we find that verbalization-based methods fail to consistently outperform likelihood-based approaches, even when applied to generative models. This limitation highlights the challenges such methods face in making fine-grained distinctions required for precise plausibility estimation.

VLMs demonstrate superior effectiveness in learning visual-related commonsense knowledge. Comparing the ComPaSS methods based on various backbones, we find VLMs exhibit particular strength in visual attribute ranking, with EVA-CLIP achieving the highest scores on CoDa (62.87), Color (51.73), and Shape (48.05), signif-

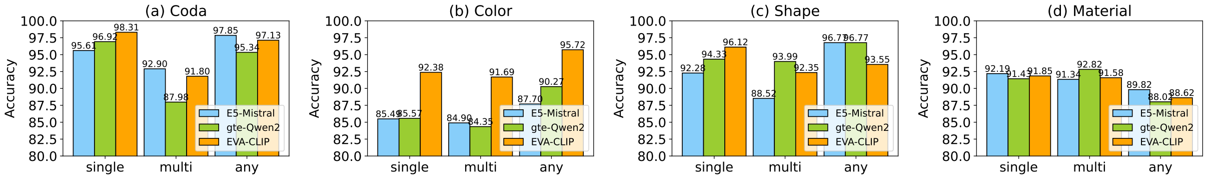


Figure 2: Binary classification accuracy of models with ComPaSS on different groups.

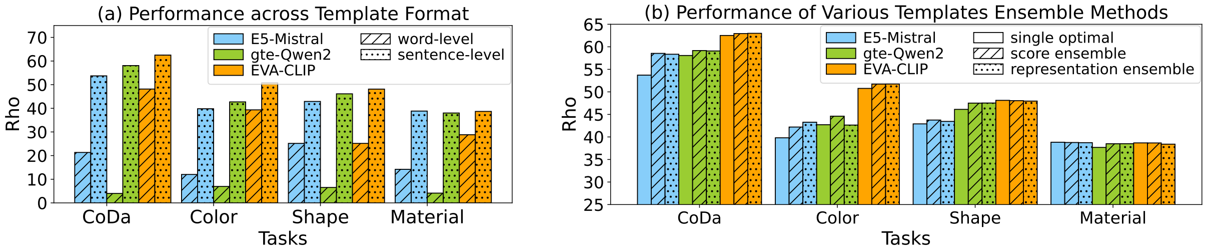


Figure 3: ComPaSS performance with different template types and template ensemble settings.

icantly outperforming even 7B parameter LLMs. This performance gap persists despite the LLMs' access to large-scale text corpora and additional parameters, underscoring the unique value of visual supervision. This performance gap highlights the limitations of text-only training, as even extensive textual data and additional parameters cannot fully compensate for the lack of visual grounding, which underscores the importance of multimodal learning for comprehensive commonsense understanding.

Discriminative approaches may offer a more parameter-efficient pathway compared to generative methods. Our experiments reveal that encoder-only models with millions of parameters like RoBERTa and CLIP-series models achieve comparable or even superior results to much larger decoder-only models (with billions of parameters) when combined with ComPaSS. This suggests that our discriminative method effectively leverages the semantic representation strengths of encoder models, which are generally more parameter-efficient than generative models. By focusing on representation-level semantics rather than token generation, ComPaSS aligns closely with the pretraining objectives of encoder models, maximizing their representation power.

The ability to discern semantic nuances in sentence representations is crucial for the performance of ComPaSS. As shown in Table 2, experiments with different RoBERTa variants reveal that applying ComPaSS to vanilla RoBERTa, which has weaker representation capabilities, leads to performance degradation. However, incorporating contrastive learning significantly improves performance, with even unsupervised contrastive training yielding substantial gains. Contrastive pre-training enables even subtle plausibility distinctions to manifest as measurable shifts in embedding space, making it essential to the performance of ComPaSS.

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
$\text{Mistral}_{\text{w/ CL}}$	94.97	86.06	91.50	**91.27**
$\text{Qwen2}_{\text{w/ CL}}$	94.71	86.79	94.04	90.42
EVA-CLIP	**95.39**	**93.29**	**94.33**	90.79

Table 3: Binary comparison accuracy on CoDa and ViComTe. The best results are highlighted in bold. All results are shown in percentage. Both Mistral and EVA-CLIP use the ComPaSS method.

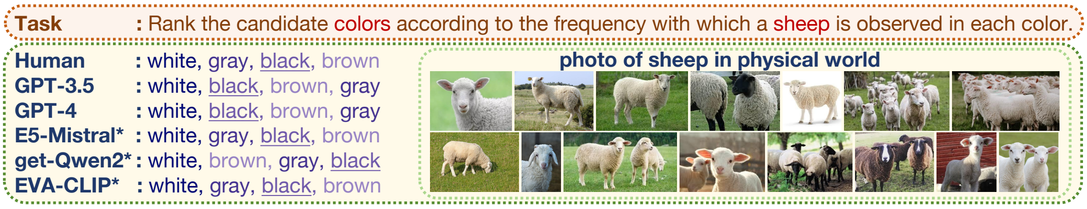
6.2 Further Analyses

6.2.1 Comparisons to Closed-source Models

We extend our evaluation to include state-of-the-art closed-source models, with results presented in Table 3. Notably, our method outperforms even GPT-4 across multiple tasks, demonstrating its effectiveness in fine-grained CSPE. This performance gap further highlights the limitations of verbalization-based approaches in capturing subtle distinctions required for precise plausibility estimation.

6.2.2 Granular Analysis of Attribute Types

We analyze binary comparison results on CoDa and ViComTe across three attribute groups: *single*: includes objects with one dominant attribute value (e.g., snow's color), *multi*: includes objects with attributes mainly distributed among the top four values (e.g., a penguin's color), and *any*: includes



Task : Rank the candidate colors according to the frequency with which a sheep is observed in each color.

Human	: white, gray, black, brown
GPT-3.5	: white, black, brown, gray
GPT-4	: white, black, brown, gray
E5-Mistral*	: white, gray, black, brown
get-Qwen2*	: white, brown, gray, black
EVA-CLIP*	: white, gray, black, brown

Figure 4: The ranking of sheep colors by humans and different models, along with corresponding images from the physical world (from Google). The '*' in the upper right represents the model with ComPaSS method.

objects with a broader attribute distribution (e.g., a T-shirt's color). As shown in Figure 3, VLMs demonstrate particular strength in the single group. This advantage primarily stems from how visual information overcomes reporting bias in textual data. For objects in the *single* category, their most common attribute is often not explicitly mentioned in text due to its widespread acceptance as common knowledge. However, these attributes are consistently and explicitly depicted in visual data. This inherent visual grounding allows VLMs to capture stereotypical attributes more effectively than text-based LLMs.

6.2.3 Effect of Template Format

We investigate the importance of sentence-level context in semantic shift measurement by comparing two approaches: *word collocation* comparison (e.g., 'penguin' and 'black penguin') and *full sentence construction* (e.g., 'There is a penguin' and 'There is a black penguin'). As shown in Figure 3(a), sentence-level inputs consistently outperform word-level comparisons for both LLMs and VLMs. This performance gap underscores the importance of complete sentence construction for ComPaSS, as sentence-level inputs better align with models' pre-training data formats.

6.2.4 Template Ensemble Methods

We investigate three template utilization strategies: The *single-optimal ensemble* approach uses the unified best-performing template, serving as an implicit ensemble. For explicit ensemble methods, *score-level ensemble* averages prediction scores across multiple templates, and *representation-level ensemble* fuses sentence representations from several templates before computing the final score. As shown in Figure 3 (b), both explicit ensemble strategies significantly improve LLM performance, with score-level ensemble showing more consistent gains. However, VLM shows limited improvement from ensemble methods, likely due to their simpler pre-training data structure. This contrast highlights LLMs' sensitivity to linguistic variations and their ability to benefit from diverse syntactic structures.

6.3 Case Study

We use the classic 'black sheep problem' to intuitively explain why ComPaSS is effective. Since 'black sheep' is an idiom, one is much more likely to mention a 'black sheep' than to specify the color of a sheep. Such reporting bias confuses the LMs that learn knowledge through probabilistic modeling. As shown in Figure 4, GPT-3.5 and GPT-4 both overestimate the probability of 'black' being the color of sheep even though sheep in black are rare. In contrast, our approach relies on semantic rather than probabilistic likelihood is able to distinguish between the linguistic meaning and the visual recognition of 'a black sheep', resulting in a more accurate estimation of the sheep's color. In addition, VLM calibrates the color distribution well by incorporating visual information.

7 Conclusion

We introduce ComPaSS, a discriminative framework for fine-grained commonsense plausibility estimation via semantic shift measurement. By leveraging the idea that plausible commonsense augmentations cause minimal semantic deviation, ComPaSS offers a generalizable approach for various tasks and model architectures. Our experiments show that discriminative methods outperform generative approaches in capturing nuanced plausibility distinctions, with ComPaSS consistently surpassing likelihood-based and verbalization-based baselines. Vision-language models also excel on visually-grounded commonsense tasks, addressing reporting bias through multimodal alignment. Finally, we emphasize the role of contrastive pre-training in improving semantic representation quality, directly enhancing plausibility estimation accuracy. Overall, ComPaSS highlights the value of utilizing semantic embeddings to extract commonsense knowledge from pre-trained models.

8 Limitations and Ethical Considerations

ComPaSS faces challenges in making absolute pointwise judgments. The method's reliance on semantic shift measurement inherently provides comparative assessments rather than definitive plausibility scores. This limitation stems from the difficulty in establishing absolute semantic distance thresholds for plausibility classification. Future work could explore calibration techniques to bridge this gap. In addition, for attribute value ranking task, 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 our method relies on LLMs and VLMs, 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

2024. Qwen2 technical report.

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*.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*.

Qi Cheng, Michael Boratko, Pranay Kumar Yelugam, Tim O'Gorman, Nalini Singh, Andrew McCallum, and Xiang Li. 2024. Every answer matters: Evaluating commonsense with probabilistic measures. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 493–506, Bangkok, Thailand. 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.

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.

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.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.

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.

Stefanie Krause and Frieder Stolzenburg. 2024. From data to commonsense reasoning: the use of large language models for explainable ai. *arXiv preprint arXiv:2407.03778*.

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.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

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.

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

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021b. 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.

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*.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. 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*.

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.

B Prompt for Sentence Transformation

The prompt we use for converting question-answer pair can be found in Figure 6.

C Prompt for Verbalization-based Method

The prompt we use for the verbalization-based method can be found in Figure **??**.

D More Experimental Results

Since not all models are compatible with all methods, we exclude the results of incompatible model-method combinations from the main text. The complete results are provided in Table 5. Notably, the results of Mistral$_{\text{w/ CL}}$ with the verbalization-based method is 0, as this model, trained via contrastive learning, has significantly lost its ability to follow instructions, preventing it from generating reasonable responses based on prompts.

```
Transform the problem into declarative sentence based on each answer with minimal modifications. Do not
introduce more information, and do not lose any information in the questions and answers.

For Example:
Question 1:
who was driving through the night, shooting blurred lights out of focus?
Answers 1:
1. person, 2. chauffeur, 3. taxi driver, 4. a person, 5. or a driver.
Sentences 1:
1. A person was driving through the night, shooting blurred lights out of focus.
2. A chauffeur was driving through the night, shooting blurred lights out of focus.
3. A taxi driver was driving through the night, shooting blurred lights out of focus.
4. A person was driving through the night, shooting blurred lights out of focus.
5. A driver was driving through the night, shooting blurred lights out of focus.
Question 2:
why would a goat eat hay in a stable?
Answers 2:
1. gain energy, 2. to fulfill hunger, 3. to get nutrition, 4. get nutrition
Sentences 2:
1. a goat eats hay in a stable to gain energy.
2. a goat eats hay in a stable to fulfill hunger.
3. a goat eats hay in a stable to get nutrition.
4. a goat eats hay in a stable to get nutrition.
Question 3:
why would an aircraft receive fuel from a cargo aircraft?
Answers 3:
1. longer flight times, 2. takeoff, 3. traveling, 4. enable travel, 5. refill fuel
Sentences 3:
1. an aircraft receives fuel from a cargo aircraft because of longer flight times.
2. an aircraft receives fuel from a cargo aircraft for takeoff.
3. an aircraft receives fuel from a cargo aircraft for traveling.
4. an aircraft receives fuel from a cargo aircraft to enable travel.
5. an aircraft receives fuel from a cargo aircraft to refill fuel.

New Task:
Question 4:
<Q>
Answers 4:
<A>
Sentences 4:
```

Figure 5: The prompt for converting question-answer pair into sentence. The blue part is the instruction, the green part is the 3-shot example, and the red part is the placeholder for the specific input.

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

The Prompt of Verbalization-based Method for Attribute Value Ranking

Sort all the <PROP>s in candidate set based on how frequently the object is observed to be each <PROP>. The higher the <PROP> is ranked, the more commonly the object is of that <PROP>. The candidate set is <PROP_LIST>. The output must be a sorted result that includes all candidate <PROP>s as in the example.

Here are some samples:
object: <OBJ0>
result: <RES0>
object: <OBJ1>
result: <RES1>
object: <OBJ2>
result: <RES2>

New Task:
object: <OBJ>
result:

The Prompt of Verbalization-based Method for Commonsense Frame Completion

Answer the question based on commonsense. Your answer should be brief. You cannot refuse to answer for any reason.

Example 1:
Question: who was driving through the night, shooting blurred lights out of focus?
Answer: person
Example 2:
Question: why would an aircraft receive fuel from a cargo aircraft?
Answer: to fly
Example 3:
Question: where's the heart-shaped hot dog and some pizza on a big tray?
Answer: restaurant

New Task:
Question: <Q>
Your answer:

Figure 6: The prompt for attribute value ranking task and commonsense frame completion task.

	Model (#Inference Parameters)	CoDa	Color	Shape	Material	CFC
	Baselines					
CSM	ACCENT (440M)	10.07	10.35	-2.10	16.99	35.04
	COMET-Atomic-2020-Bart (440M)	22.91	26.98	40.44	25.72	-
	VERA-T5-XXL (5B)	58.93	45.08	30.31	33.51	45.81
LM	RoBERTa+likelihood (355M)	24.37	33.63	36.12	24.23	42.46
	RoBERTa$_{\text{w/ CL}}$+likelihood (355M)	23.36	31.51	26.69	22.23	38.03
	Mistral+verbal. (7B)	46.64	38.63	30.46	36.34	32.06
	Mistral+likelihood (7B)	51.30	34.31	26.70	37.03	47.98
	Mistral$_{\text{w/ CL}}$+verbal. (7B)	0.0	0.0	0.0	0.0	0.0
	Mistral$_{\text{w/ CL}}$+likelihood (7B)	25.70	4.72	18.81	5.96	35.46
	Qwen2+verbal. (7B)	57.40	41.59	38.3	36.76	29.32
	Qwen2+likelihood (7B)	50.25	40.99	32.52	37.13	45.10
	Qwen2$_{\text{w/ CL}}$+verbal. (7B)	11.12	15.28	-24.21	0.45	21.39
	Qwen2$_{\text{w/ CL}}$+likelihood (7B)	49.65	41.75	32.8	37.3	43.00
	ComPASS					
LM	RoBERTa$_{\text{w/ CL}}$ (355M)	44.59	38.92	42.92	33.55	44.46
	Mistral$_{\text{w/ CL}}$ (7B)	58.54	42.20	43.75	**38.77**	**49.01**
	Qwen2$_{\text{w/ CL}}$ (7B)	59.16	44.61	47.51	38.49	46.41
VLM	CLIP (124M)	58.10	45.55	45.82	33.56	35.13
	EVA-CLIP (695M)	**62.87**	**51.73**	**48.05**	38.67	41.46

Table 5: Spearman's rank correlation coefficient ρ between the predicted ranks of candidates and their ground-truth on CoDa, ViComTe (Color, Shape, and Material), and CFC, shown in percentage. The **best** and second best results are highlighted in bold and underlined, respectively. '+likelihood' indicates using the likelihood-based method and '+verbal.' indicates using the verbalization-based method.