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 varying input formats and commonsense knowledge levels based on 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 of large language models (LLMs), ensuring that generated content is not only fluent but also consistent with commonsense world knowledge remains challenging (Marcus, 2020; Chen et al., 2023). More broadly, prior work has shown that LMs may exhibit inconsistencies and gaps between linguistic competence and grounded world understanding (Elazar et al., 2021; Mahowald et al., 2024). Reliable commonsense plausibility evaluation is therefore important for identifying implausible or inconsistent generations and for providing quantitative feedback to improve LLM outputs (Tian et al., 2023).

Motivated by this need, we study commonsense plausibility estimation (CSPE), which assesses how plausible a piece of information is with respect to commonsense knowledge. Prior work has operationalized commonsense plausibility in different ways, including scoring individual statements, distinguishing commonsensical from nonsensical statements, choosing the more plausible alternative, and modeling graded plausibility over candidate attributes or answers (Liu et al., 2023; Wang et al., 2020; Roemmele et al., 2011; Paik et al., 2021; Zhang et al., 2022; Cheng et al., 2024). Unlike standard commonsense QA tasks such as CommonsenseQA (Talmor et al., 2019), PIQA (Bisk et al., 2020), and SocialIQA (Sap et al., 2019b), which typically require selecting a single best answer, many CSPE settings focus on plausibility itself and allow multiple candidates to be reasonable to different degrees. In this paper, we focus on training-free comparative CSPE, where a model scores candidates under the same context and ranks them by relative plausibility.

Existing training-free approaches to CSPE mainly adopt a generative perspective. Likelihood-based methods (Trinh and Le, 2018; Tamborrino et al., 2020; Holtzman et al., 2021) use token prediction probabilities as plausibility indicators, assuming that commonsense-consistent sentences tend to receive higher likelihoods. Verbalization-based methods (Brown et al., 2020; Krause and Stolzenburg, 2024) prompt LMs to judge or rank plausibility through natural language responses. However, such generative approaches can be suboptimal for fine-grained comparative CSPE, where the goal is not only to assess whether a sentence is likely, but

also to discriminate subtle plausibility differences among multiple partially valid candidates.

We instead adopt a discriminative perspective. 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 observation, we propose ComPaSS, a method that measures **Com**monsense **Pla**usibility through **S**emantic **S**hifts introduced by adding candidate information to a sentence. Plausible additions should induce small semantic shifts, whereas implausible ones should cause larger deviations. For example, adding 'black' to 'There is a penguin' introduces only a minor shift, as it matches penguins' typical appearance; by contrast, adding 'green' produces a larger shift because it introduces an atypical attribute. ComPaSS quantifies this shift by computing the embedding similarity between an anchor sentence and its candidate-augmented counterpart.

The effectiveness of ComPaSS depends on the quality of semantic representations. We therefore examine two factors: *modality* and *contrastive learning*. First, language models may suffer from reporting bias (Gordon and Durme, 2013), where obvious commonsense facts, such as typical object colors or shapes, are rarely stated explicitly in text. Vision-language models (VLMs), in contrast, can provide visual grounding for attributes such as color, shape, and material (Paik et al., 2021; Zhang et al., 2022). Second, contrastive pre-training can produce more discriminative sentence representations, making fine-grained semantic shifts easier to measure. We therefore evaluate ComPaSS with both LM and VLM backbones, including models with and without contrastive pre-training.

We evaluate ComPaSS on two types of fine-grained CSPE tasks. The first is attribute value ranking, using CoDa (Paik et al., 2021) and ViComTe (Zhang et al., 2022), where models rank candidate values such as colors, shapes, and materials for common objects. The second is commonsense frame completion (Cheng et al., 2024), where models rank plausible completions for free-form open-ended questions. For example, given "Where are farmers with newly harvested crops?", a plausible completion such as 'farm' should be ranked above a less plausible one such as 'truck'. In this free-form setting, ComPaSS first converts question-answer pairs into declarative sentence pairs before applying semantic-shift scoring. Together, these tasks evaluate ComPaSS across structured and free-form inputs, as well as across visually grounded attributes and broader everyday commonsense.

Our experiments show three main findings. First, ComPaSS generally outperforms likelihood-based and verbalization-based baselines in our evaluated fine-grained ranking settings, suggesting that semantic-shift scoring is a strong alternative to generative plausibility estimation. Second, VLMs combined with ComPaSS show particular advantages on visually grounded attributes such as color and shape, highlighting the value of visual grounding for commonsense plausibility estimation. Third, contrastive pre-training substantially improves ComPaSS, indicating that discriminative sentence representations are crucial for reliable semantic-shift measurement.

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). Recent studies on event and semantic plausibility further show that LM probabilities can capture plausibility distinctions to some extent, such as the gap between impossible and unlikely events and semantic plausibility in base and instruction-tuned LMs (Kauf et al., 2023, 2024). For LLMs with instruction-following capability, they can be prompted to judge whether an input is commonsense-consistent (Zhao et al., 2024). Beyond judging plausibility, some methods (Jung et al., 2022; Tafjord et al., 2022) evaluate hypotheses by scoring entailment paths generated by LLMs, i.e., reasoning chains justifying 'reasonable' or 'unreasonable' conclusions, and selecting the final prediction from the highest-scoring path. VERA (Liu et al., 2023) adopts a discriminative approach by fine-tuning a classification head on model representations using ~7 million commonsense statements. In contrast, ComPaSS leverages internal knowledge from a discriminative perspective, but estimates plausibility through representation-level semantic shifts without additional training.

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 typi-

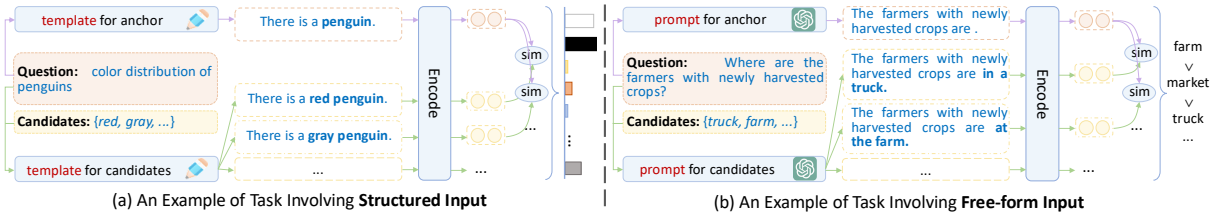

Figure 1: How ComPaSS works on different tasks.

cal 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., 2019a; 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

Commonsense plausibility estimation (CSPE) assesses the plausibility of candidate information under a given context; here we consider comparative CSPE, where candidates associated with the same context are ranked by their relative plausibility.

Formally, given an input instance $x_i = (c; a_i^c)$ consisting of a context c and candidate information $a_i^c \in A^c$, 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 ranking score set $\mathcal{P}^c = \{p_1^c, p_2^c, ..., p_K^c\}$ for all candidates. In practice, a model may score each context-candidate pair (c, a_i^c) independently, and the scores of all candidates under the same context are then collected to form $\mathcal{P}^c$. Each $p_i^c \in \mathbb{R}$ indicates the relative 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 human-annotated plausibility or preference score of a_i^c. Performance is measured by the correlation between $\mathcal{P}^c$ and $\mathcal{G}^c$.

We consider two instantiations. In *attribute value ranking*, $c = (o, \text{has property } p)$, where o is a common object and p is a property, and each candidate a_i^c is an attribute value for p. In *commonsense frame completion*, $c = q$ is a free-form question, and each a_i^c is a possible answer.

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 *structured triplet inputs*, we employ template-based construction, where z represents a pre-defined template (see Appendix B) and $f(\cdot)$ represents applying this template to generate a sentence. In contrast, for tasks involving *free-form question-answer pairs as input*, we query GPT-4 (Achiam et al., 2023) to generate contextually coherent sentences, where z denotes the prompt (see Appendix C) and $f(\cdot)$ represents querying GPT-4 using the prompt. Since

questions cannot be directly converted into coherent statements, we use a blank space as a placeholder when constructing anchor sentences. Such an adaptive sentence construction method enables ComPaSS to be applicable to different input forms.

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 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 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 everyday 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 is derived from Visual Genome (Krishna et al., 2017). It offers attribute

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

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

value distributions across broader 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.

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, which should align 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 rank correlation between predicted and ground-truth plausibility orderings. 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.

Accuracy: CoDa and ViComTe also include binary comparison tasks where each object is paired with two attribute values, with one more plausible than the other. Models need to rank the more plausible value higher. Accuracy quantifies the success rate of these binary selections. This metric is suitable for cross-attribute comparisons as it is unaffected by variations in the number of candidates, unlike 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 evaluate both base models and their contrastive learning pre-trained variants: RoBERTa-Large (Liu et al., 2021b) (RoBERTa) is a widely-used encoder-only LM with fewer parameters. Mistral-7B-Instruct (Jiang et al., 2023) (Mistral) and Qwen2-7B-instruct (qwe, 2024) (Qwen2) are two decoder-only LLMs with strong instruction-following capabilities. We also evaluate their **contrastive learning pre-trained** 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$_{\text{w/ CL}}$). Please note that all contrastive learning procedures are pre-training stage optimizations unrelated to our task. We directly use their released checkpoints without task-specific fine-tuning.

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).

5.3.2 Baselines

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 to directly estimate their plausibility.

Language models (LMs): We evaluate all open-source LMs used as the backbone of ComPaSS with two methods. For the *likelihood based* method, the plausibility of a sentence is proportional to the normalized probability of predicting each token sequentially. For the *verbalization based* method, pre-trained LMs are prompted in natural language (see Appendix D) to rank candidates based on 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 multimodal technology with superior capabilities.

5.4 Implementation Details

All experiments are carried out in a zero-shot or in-context few-shot setting. Closed-source models are accessed via official APIs, while open-source implementations run on a single NVIDIA A800 80G GPU. For ACCENT, the beam number is 10

[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
CSM	COMET-Atomic (440M)	22.91	26.98	40.44	25.72	-
CSM	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
LM	RoBERTa$_{w/\ CL}$+likelihood (355M)	23.36	31.51	26.69	22.23	38.03
LM	Mistral+verbal. (7B)	46.64	38.63	30.46	36.34	32.06
LM	Mistral+likelihood (7B)	51.30	34.31	26.70	37.03	47.98
LM	Qwen2+verbal. (7B)	57.40	41.59	38.30	36.76	29.32
LM	Qwen2+likelihood (7B)	50.25	40.99	32.52	37.13	45.10
LM	Qwen2$_{w/\ CL}$+likelihood (7B)	49.65	41.75	32.80	37.30	43.00
	ComPaSS					
LM	RoBERTa$_{w/\ CL}$ (355M)	44.59	38.92	42.92	33.55	44.46
LM	Mistral$_{w/\ CL}$ (7B)	58.54	42.20	43.75	**38.77**	**49.01**
LM	Qwen2$_{w/\ CL}$ (7B)	<u>59.16</u>	44.61	<u>47.51</u>	38.49	46.41
VLM	CLIP (124M)	58.10	<u>45.55</u>	45.82	33.56	35.13
VLM	EVA-CLIP (695M)	**62.87**	**51.73**	**48.05**	<u>38.67</u>	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 <u>second best</u> results are highlighted in bold and underlined, respectively. '+verbal.' indicates using the verbalization-based method.

as the official setting. For CFC, where the original candidate answers are short fragments rather than standalone sentences, we apply GPT-4 sentence transformation as a shared preprocessing step for all methods. When using the verbalization-based method on CFC, we sample the model 100 times for each question with a temperature of 0.7, and cluster answers following the official protocol.

6 Results and Analysis

6.1 Overall Results

The overall experimental results are presented in Table 1, which reveals several key findings:

ComPaSS achieves the best performance compared to baselines across both structured triplets (attribute ranking) and free-form text (CFC) inputs. This demonstrates its robustness to diverse input formats without relying on task-specific templates. 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, whereas ComPaSS succeeds by unifying semantic shift measurement across both templated and non-templated scenarios.

Model / Method	CoDa	Color	Shape	Material
Likelihood-based scoring				
RoBERTa (vanilla)	24.37	33.63	36.12	24.23
ComPaSS				
RoBERTa (vanilla)	24.63	22.68	26.77	19.93
RoBERTa (unsup. CL)	32.67	32.00	42.18	31.12
RoBERTa (sup. CL)	44.59	38.92	42.92	33.55

Table 2: Effect of sentence representation quality on ComPaSS with RoBERTa variants. The likelihood row uses vanilla RoBERTa as a reference baseline. Values are Spearman's ρ (%).

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), significantly outperforming even 7B parameter LLMs.

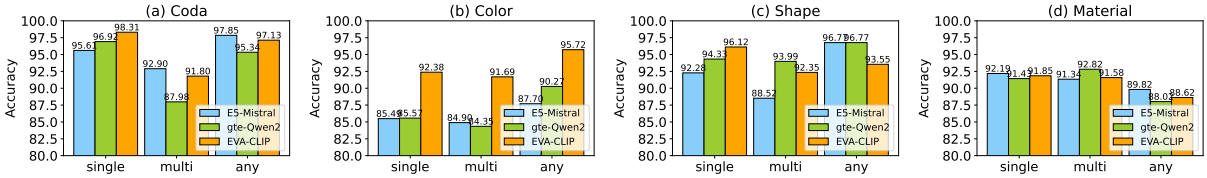

Figure 2: Binary classification accuracy of models with ComPaSS on different groups.

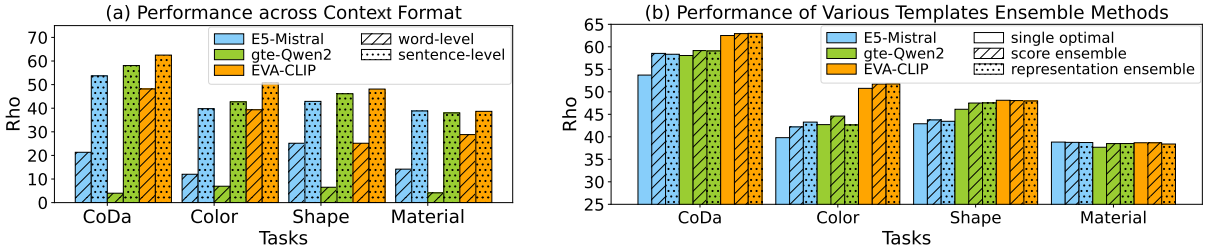

Figure 3: ComPaSS performance with different context formats and ensemble settings.

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 pre-training objectives of encoder models, maximizing their representation power.

The ability to discern semantic nuances in sentence representations is crucial for ComPaSS. As shown in Table 2, experiments with different RoBERTa variants reveal that applying ComPaSS to vanilla RoBERTa leads to performance degradation due to its weaker representation capabilities. However, incorporating contrastive learning (even via unsupervised training) significantly improves performance by enabling subtle plausibility distinctions to manifest as measurable embedding space shifts. Crucially, ComPaSS does not require cus-

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.

tom contrastive pre-training. It directly leverages contrastively pre-trained SOTA embedding models, enabling continuous performance gains from evolving embedding techniques without task-specific fine-tuning or architectural modifications.

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

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 2, VLMs demonstrate particular strength in the single group. We attribute this advantage to visual grounding overcoming textual reporting bias: stereotypical attributes are rarely explicitly stated in text due to their commonsense nature, creating a reporting bias in language data. However, these attributes are consistently and explicitly depicted in images, enabling VLMs to overcome linguistic omissions. This finding demonstrates that visual grounding serves as a critical compensator for missing commonsense in text-based training.

6.2.3 Effect of Context 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

For the template-based method, we investigate three ensemble 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 further improve LLM performance, with the score-level ensemble showing more consistent gains. However, VLM shows limited improvement from ensemble methods, likely due to its 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

ComPaSS is better suited for relative ranking than for absolute plausibility judgment, as its semantic-shift measurements are inherently comparative. Therefore, its scores should be interpreted within a given context rather than as calibrated plausibility probabilities across contexts. Future work could explore calibration methods for defining more reliable plausibility thresholds.

9 Ethical Considerations

As ComPaSS relies on LLMs and VLMs, it may inherit biases in their training data and representation spaces. Such biases, including societal stereotypes or skewed attribute associations, could affect plausibility rankings and may be perpetuated or amplified by semantic-shift scoring. Addressing these biases is an important direction 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*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7432–7439.

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.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

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.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4463–4473.

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*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

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.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. Semeval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 307–321.

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 AI Assistants

AI tools were used solely to enhance the clarity and fluency of the manuscript. They were not used to generate research ideas, design experiments, analyze data, interpret results, or draw conclusions. All such tasks were conducted and independently verified by the authors.

B Templates for Sentence Construction

The templates we used to construct anchor sentences and candidate sentences of different property are shown in Table 4.

C Prompt for Sentence Transformation

The prompt we use for converting question-answer pair can be found in Figure 5. For the Commonsense Frame Completion (CFC) task, answers with similar semantics (e.g., "person" vs. "a person") will be further grouped into equivalence clusters during evaluation rather than being considered as individual answers. Following the dataset's official protocol, each question is asked multiple times to estimate the sampling probability of the model as accurately as possible, and different expressions of the same type of answer are allowed to avoid the influence of vocabulary selection on the model.

D Prompt for Verbalization-based Method

The prompt we use for the verbalization-based method can be found in Figure 6. We describe the parsing and post-processing procedure below.

Parsing and post-processing. For verbalization-based ranking, we prompt the model to output a JSON-like dictionary mapping each context object or question to an ordered list of candidates. We

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.

parse the output by normalizing generated strings (e.g., lowercasing) and matching them to the predefined candidate set. Malformed, incomplete, or unparsable outputs are retried with a format-reminder prompt up to 10 times. After retries, duplicated candidates are kept only at their first occurrence, unrecognized candidates are ignored, and missing candidates are appended to the end in their original order. The same protocol is applied to all verbalization-based baselines.

E More Experimental Results

Since some checkpoints are less compatible with certain inference methods, we exclude these additional model-method combinations from the main text for readability and report the complete results in Table 5. Notably, $\text{Mistral}_{\text{w/ CL}}$ yields 0.0 with the verbalization-based method across all tasks, because this contrastively trained checkpoint is optimized for sentence embedding quality and no longer reliably follows ranking prompts in our setting.

F Computational Complexity

For each context with K candidates, ComPaSS encodes one anchor sentence and K candidate sentences. The anchor representation is computed only once and shared across candidates, while candidate

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.

sentences can be encoded in parallel within a batch. Thus, the encoding cost scales linearly with the candidate set size, i.e., $O(K)$, and the subsequent similarity computation costs $O(Kd)$, where d is the embedding dimension. Compared with verbalization-based ranking, ComPaSS avoids autoregressive generation and output parsing; compared with likelihood-based scoring, it avoids token-level probability computation for each candidate sentence.

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.

	Model (#Inference Parameters)	CoDa	Color	Shape	Material	CFC
	Baselines					
CSM	ACCENT (440M)	10.07	10.35	-2.10	16.99	35.04
CSM	COMET-Atomic-2020-Bart (440M)	22.91	26.98	40.44	25.72	-
CSM	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
LM	RoBERTa$_{\text{w/ CL}}$+likelihood (355M)	23.36	31.51	26.69	22.23	38.03
LM	Mistral+verbal. (7B)	46.64	38.63	30.46	36.34	32.06
LM	Mistral+likelihood (7B)	51.30	34.31	26.70	37.03	47.98
LM	Mistral$_{\text{w/ CL}}$+verbal. (7B)	0.0	0.0	0.0	0.0	0.0
LM	Mistral$_{\text{w/ CL}}$+likelihood (7B)	25.70	4.72	18.81	5.96	35.46
LM	Qwen2+verbal. (7B)	57.40	41.59	38.3	36.76	29.32
LM	Qwen2+likelihood (7B)	50.25	40.99	32.52	37.13	45.10
LM	Qwen2$_{\text{w/ CL}}$+verbal. (7B)	11.12	15.28	-24.21	0.45	21.39
LM	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
LM	Mistral$_{\text{w/ CL}}$ (7B)	58.54	42.20	43.75	**38.77**	**49.01**
LM	Qwen2$_{\text{w/ CL}}$ (7B)	<u>59.16</u>	44.61	<u>47.51</u>	38.49	46.41
VLM	CLIP (124M)	58.10	<u>45.55</u>	45.82	33.56	35.13
VLM	EVA-CLIP (695M)	**62.87**	**51.73**	**48.05**	<u>38.67</u>	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 <u>second best</u> results are highlighted in bold and underlined, respectively. '+likelihood' indicates using the likelihood-based method and '+verbal.' indicates using the verbalization-based method.