Evaluation of event plausibility recognition in Large (Vision)-Language Models

Maria Cassese*
Università di Pisa

Alessandro Bondielli**
Università di Pisa

Alessandro Lenci†
Università di Pisa

Transformer-based Language Models (LMs) achieve outstanding performances in various tasks but still exhibit limitations in recognizing common world events (GEK), particularly when they require referential information or real-world experience. Assuming that visual knowledge in vision-language models (VLMs) provides additional referential information, this paper tests their ability to leverage implicit event knowledge to acquire robust and generalizable representations of agent-patient interactions, assessing their capacity to distinguish between plausible and implausible events. The analysis was conducted on models of varying sizes and architectures.

In the evaluation, the performance of unimodal and multimodal models of various sizes was compared using the task of recognizing the plausibility of minimal sentence pairs. Our analysis suggests several findings: 1) decoder-only models tend to outperform encoder-only ones; 2) the model size has a minor impact: although larger models perform better in absolute terms, the differences between 7B and 13B parameter models are not significant for this particular task; 3) while smaller encoder-only VLMs consistently fall short of their LLM counterpart, larger ones have similar or slightly superior performance; 4) all models have lower performance on the more challenging sentences; 5) adding corresponding images to the textual stimuli affects the accuracy levels of some models. These findings open avenues for further analyses of the inner workings of VLMs and their ability to model event knowledge with and without visual inputs.

1. Introduction

Human linguistic knowledge develops alongside daily interactions with entities and objects of the external world. When we read or hear a word, along with its core meaning, information related to verb selection preferences or typical event participants is activated. This type of knowledge allows us to immediately recognize that the event *"A cop arrested a thief"* is plausible, while *"A thief arrested a cop"* is unlikely, and *"A stone arrested a thief"* is impossible. Expectations regarding the prosecution of a sentence are dynamically updated. For example, imagine reading the sequence *"The boss"*. Depend-

* Department of Computer Science, University of Pisa, 3 Largo Bruno Pontecorvo, Pisa, 56127, Italy. E-mail: maria.cassese@phd.unipi.it
** University of Pisa, 3 Largo Bruno Pontecorvo, Pisa, 56127, Italy. E-mail: alessandro.bondielli@unipi.it
† CoLing Lab, Department of Philology, Literature, and Linguistics, University of Pisa, 36 S. Maria St, Pisa, I-56126, Italy. E-mail: alessandro.lenci@unipi.it

ing on whether the next verb is *"fired"* or *"approved"*, our minds envision two different scenarios and dynamically adjust these expectations (Matsuki et al. 2011).

The ability to predict the likelihood of events derives from the direct and indirect experience of the world and its events, through which humans form an abstract prototypical representation called generalized event knowledge (GEK) (McRae and Matsuki 2009), based on which we classify new combinations of topics as typical or atypical.

This representation is inherently multimodal, constructed by processing textual, visual, and auditory information (Baltrušaitis, Ahuja, and Morency 2018).

The acquisition of GEK in language models trained on large text corpora is limited by the so-called *reporting bias*. This bias implies that in both written and oral communication, uncommon events are more frequently mentioned than common ones (Gordon and Van Durme 2013). Transformer-based language models (Vaswani et al. 2017) have been shown to only partially acquire GEK compared to humans (Pedinotti et al. 2021; Kauf et al. 2022).

In our previous work (Cassese, Bondielli, and Lenci 2023), the initial assumption was that the complementary information possessed by vision-language models would improve their capability to represent the plausibility of events (Bruni et al. 2012). Contrary to expectations, analyses showed that vision-language encoder-only models such as VisualBERT and FLAVA perform comparably to BERT and RoBERTa. Specifically, it was found that the VLMs did not accurately interpret compositional information when the word order in a sentence was altered, instead behaving like bag-of-words models. Given these findings, the present study examines the impact of architecture and model size by utilizing decoder-only models with billions of parameters. The new models evaluated are LLAVA-Mistral, LLAVA-Vicuna, and their corresponding unimodal versions, Mistral-Instruct and Vicuna.

All the models sampled are open-weight models taken from Huggingface [1] considering the following aspects: first, we have taken into account encoder-only vs. decoder-only models; second, we considered LLMs with a corresponding open-weight VLM based on it; third, we considered the parameter size of decoder only LLMs and VLMs.

The chosen models were evaluated on three datasets consisting of pairs of sentences distinguished by plausibility and argument animacy (in Table 1, detailed data regarding the utilized datasets are presented). As shown in fig. 1, the employed datasets have three different levels of difficulty. The EventsAdapt dataset consists of two subsets: i) the first subset contains pairs of sentences where the implausible sentence is impossible (EventsAdapt Animate Inanimate); ii) the second subset features pairs of sentences where the implausible sentence is unlikely but not impossible (EventsAdapt Animate Animate); iii) lastly, the DTFit dataset includes pairs of sentences that exhibit varying degrees of prototypicality, despite both being likely.

As previously done in (Cassese, Bondielli, and Lenci 2023), the capabilities of the models were also evaluated on subsets of concrete and abstract sentences to assess the performance of VLMs with respect to the concreteness dimension. Finally, the VLMs were tested on a smaller dataset comprising both images and texts to assess the impact of incorporating visual input on performance.

Our analysis reveals the following. First, we see that decoder-only models generally outperform encoder-only ones (in this comparison, the models' size has not been taken into account). Second, model size is relevant only partially. While larger models perform better in absolute terms, the differences between 7B and 13B parameter models

1 https://huggingface.co

Table 1
Datasets sentence examples: the animacy of the event participants is specified in parenthesis (AnIn = animate agent, inanimate patient; AnAn = animate agent, animate patient.

Dataset	Plausible?	Concreteness	Example
EventsAdapt			
(AnAn, unlikely)	Yes	Concrete	The nanny tutored the boy.
	No		The boy tutored the nanny.
	Yes	Abstract	The boss fired the worker.
	No		The worker fired the boss.
EventsAdapt			
(AnIn, impossible)	Yes	Concrete	The secretary organized the desk.
	No		The desk organized the secretary.
	Yes	Abstract	The raider caught the illness.
	No		The illness caught the raider.
DTFit			
(AnIn, unlikely)	Yes	Concrete	The barber cut the hair.
	No		The barber cut the cake.
	Yes	Abstract	The priest promised the salvation.
	No		The priest promised the promotion.
EventsRev			
(AnAn, unlikely)	Yes	Concrete	The cop is arresting the criminal.
	No		The criminal is arresting the cop.

appear to be negligible on the chosen task. Third, while smaller encoder-only VLMs are consistently outperformed by their LLM counterparts, larger ones perform on par or slightly better than LLMs on the task. Fourth, we show that all models continue to exhibit lower performance on the more challenging sentences. These findings are partly in line with previous works (Cassese, Bondielli, and Lenci 2023), but provide further indications on the behavior of VLMs and highlight the increased effectiveness of state-of-the-art VLMs in properly modeling linguistic knowledge. Finally, we consider two hypotheses for understanding whether including images is beneficial for the models' ability to understand and process their input.

The structure of this paper is the following: In Section 2 there is a review of related work. Then, Section 3 describes the datasets (Sec. 3.1), the tested models (Sec. 3.2), and the evaluation procedure (Sec. 3.3). The results are presented and discussed in Section 4, and, finally, the conclusions in Section 5.

2. Related Work

The idea of creating language and vision models originates from cognitive linguists' theories emphasizing the importance of mental conceptualization and the anchoring of meaning to extralinguistic entities (Miller and Charles 1991). George Miller and Walter Charles synthesized these ideas with their contextual hypothesis, based on linguistic use and inspired by Wittgenstein's later work, asserting that language's function can be understood through its use in context (Wittgenstein 1953). In *"Contextual Correlates of Semantic Similarities,"* they define context as the set of conditions that govern word use, suggesting that a word's contextual representation is an abstract cognitive structure derived from its encounters in linguistic contexts. This broader definition of context includes both linguistic and extralinguistic information (Morris 1938). Marconi later argues that using a word involves not only knowing its distributional constraints but

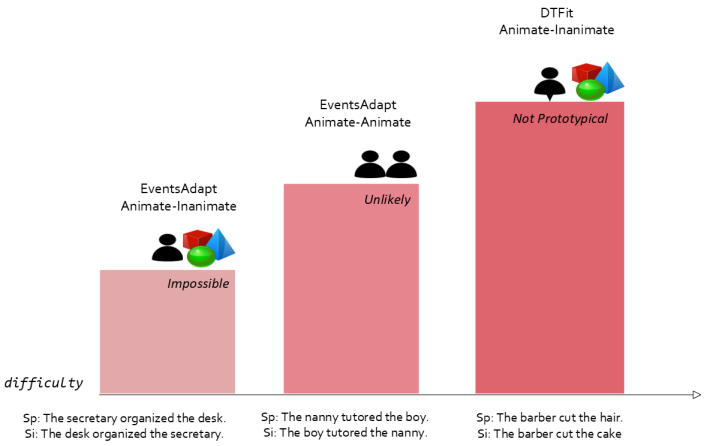

Figure 1
Datasets are separated by difficulty level. The EventsAdapt Animate-Inanimate dataset contains pairs of sentences where the implausible sentence is impossible, making the difference between the two variants easier to recognize. The EventsAdapt Animate-Animate dataset has pairs of sentences with both agent and patient being animate, so the implausible sentence is unlikely but not impossible. Finally, in the DTFit Animate-Inanimate dataset, the implausible sentence is created by replacing the typical patient with an atypical one, considering the context of the sentence.

also understanding the properties and relationships of the objects it refers to, requiring both inferential and referential competence (Marconi 1997). While inferential knowledge can be partially extracted with distributional representations from textual input, it is more difficult to do the same for referential knowledge. The use of visual information in models is also justified by the need to introduce visual grounding that enhances their knowledge and experience of the real world (Harnad 1990). In machine learning, around 2010, the first vision-language models that unified the processing of images and text were developed, based on the assumption that a common latent structure exists between words and their visual representation (Feng and Lapata 2010).

The results obtained from the earliest models have strongly supported the multimodal hypothesis, enhancing the semantic representation of concrete concepts and properties (Bruni et al. 2012). However, they have proven to be less effective in representing verbs, adjectives, and abstract concepts (Shekhar et al. 2017). Another more recent work that studied the alignment between multimodal representations and human semantic intuitions is that of Pezzelle et al. (Pezzelle, Takmaz, and Fernández 2021). In this work, static word representations were obtained from the contextualized representations learned by the models. In this case, it was observed that multimodal representations are advantageous for the representation of concrete words, but not for abstract ones.

As model training techniques for both visual (LeCun et al. 1998) and textual (Sherstinsky 2020) data advance, along with the widespread use of transfer learning methods, model capabilities have also advanced rapidly. The first visual transformer (Dosovitskiy et al. 2021) arises from the intuition that an image can be partitioned into sections and represented similarly to words, paving the way for the rapid development of multimodal Transformer models.

VLMs have demonstrated high performance on various multimodal tasks; however, they still have limitations in modeling natural language understanding, especially in cases where they need to encode abstract symbolic structures, such as in tasks where it is necessary to combine units of meaning into larger units, a capability that language models lack (Pavlick 2023). The visual grounding in VLMs is not sufficient to overcome this limitation (Thrush et al. 2022). These models fail to recognize relationships, lack sensitivity to the order of components in sentences or images, and make errors in linking objects to their attributes. Standard retrieval tasks, such as image-text matching, are successfully executed using shortcut strategies (Geirhos et al. 2020) that do not incorporate composition and order information. This limitation is highlighted in the analysis conducted in (Yuksekgonul et al. 2023), where the models' relational and attributional knowledge was assessed by identifying and isolating detailed subtypes of compositions.

Building upon existing literature, this study contributes to elucidating the limitations of VLMs in capturing compositionality while exploring the potential of LLMs in enhancing semantic representation. To this end, the performance of VLMs and LLMs is evaluated in a specific linguistic acceptability task: event plausibility recognition.

3. Experiments

We present a set of experiments spanning different datasets and models to assess whether models are able to differentiate between plausible and implausible inputs. We consider three different datasets composed of sentence pairs, where one item of the pair is a plausible sentence and the other is not. In addition to this, one of the datasets includes also visual stimuli depicting both sentences in the pair. We describe the datasets and their degree of implausibility in Section 3.1. We consider a set of models that are representative of widespread open-weight model classes. We consider: i) encoder-only and decoder-only models; ii) language-only models and vision-language ones; iii) smaller and larger models in terms of parameters. Considered models are described in Section 3.2. We evaluate all models via intrinsic metrics, i.e. Pseudo Log Likelihood (PLL) for encoder-only models, and Perplexity (PPL) for decoder-only ones (see Sec. 3.3). First, we feed the model sentence pairs from a dataset and assess whether the PLL/PPL of the plausible sentence is lower than that of the implausible one. We evaluate the results in light of the differences in plausibility of the various datasets. We evaluate both LMs and VLMs. Note that in this case VLMs are fed only the textual inputs. Second, we evaluate how including also corresponding visual stimuli affects the different VLMs' ability to recognize plausibility.

3.1 Dataset

For our experiments, we used three datasets that have also been tested in (Kauf et al. 2022). All the datasets consist of pairs of sentences categorized by plausibility, containing a plausible sentence S_p and its corresponding implausible version S_i, obtained by altering S_p. The sentences are minimal sequences composed of an agent, verb, and patient, describing transitive events. Each sentence pair was labeled according to human plausibility judgment rated on a Likert scale from 1 to 7.

The datasets differ from each other in a few aspects:

EventsAdapt (Fedorenko et al. 2020). It contains 257 pairs of sentences, each of which includes (i) a plausible sentence that describes a transitive event in the past tense

(e.g., *"The secretary organized the desk"*) and (ii) the implausible counterpart, obtained by reversing the order of the noun phrases (e.g., *"The desk organized the secretary"*). In the dataset, two distinct subsets of data are identified based on the animacy of the patient. In the subset denoted as $EventsAdapt_{AN-AN}$, both the agent and the patient are animate (e.g., *"The nanny tutored the boy"*), while in the subset labeled as $EventsAdapt_{AN-IN}$, the agent of the plausible sentence is animate while the patient is inanimate (e.g., *"The raider caught the illness"*). In this second case, reversing the order of the agent and patient results in an impossible sentence, violating the verb's selection preferences (e.g., *"The illness caught the raider"*) (Table 1).

DTFit (Vassallo et al. 2018). It comprises 395 pairs of sentences describing a transitive event in the past tense, distinguished by the typicality of the patient. Given the plausible sentence, the implausible (i.e., atypical) one is obtained by replacing the patient with a filler less typical for the sentence context (e.g., *"The barber cut the hair"* vs *"The barber cut the cake"*). In this case, we have always an animate agent that interacts with an inanimate patient. The changing factor is the degree of typicality of the patient with respect to the verb-agent pair. For instance, the cake is a typical patient for the cutting event but is less typical than cutting the hair if the agent is a barber. Consequently, typicality is defined by the word content rather than by the word order (Table 1).

EventsRev (Ivanova et al. 2021). It consists of 38 concrete sentences depicting transitive events in the present progressive tense. Within this dataset, implausible sentences are created by reversing the animate noun phrases (e.g., *"The cat is chasing the mouse"* vs. *"The mouse is chasing the cat"*)(Table 1).

Further analysis was conducted by dividing the sentences of DTFit and EventsAdapt into subsets based on their level of concreteness, resulting in the subsets $DTFit^{abstr}$, $DTFit^{concr}$ $EventsAdapt^{abstr}$ e $EventsAdapt^{concr}$. To achieve this categorization, the concreteness level of the sentence was annotated progressively, starting with the individual components (verb, agent, and patient), and subsequently determining the overall concreteness of the sentence.

3.2 Models

For model selection, we build from previous work (Cassese, Bondielli, and Lenci 2023) and include several new models. Specifically, we included two widely popular decoder-only LLMs, namely Mistral and Vicuna and their VLM counterparts based on the LLAVA architecture (Liu et al. 2023). This allows us to understand whether current trends in both LLMs and VLMs provide improvements over older encoder-only models. In addition to this, we considered two size variants of the LLAVA-Vicuna model, i.e. the 7B (7 Billions parameters) and 13B (13 Billions parameters) variants, to investigate the role of model size in this kind of task.

The models used in the experiments are described below:

VisualBERT (Li et al. 2019). VisualBERT takes as input a text and a set of regions identified by an R-CNN network in the input image, aligning the text and image regions using self-attention. The visual and textual embeddings are then concatenated and fed into a stack of transformer encoder layers, resulting in a joint representation. A BERT model is employed to generate the textual representation. The VisualBERT model has been tested on four tasks, including Visual Commonsense Reasoning (VCR) (Zellers et al. 2019).

FLAVA (Singh et al. 2021). The FLAVA model was developed to excel in both visual and textual tasks independently, as well as in vision-language tasks. To this end, a set of unimodal and multimodal pretraining objectives were defined, and the model was tested on 35 tasks. The model independently extracts visual and textual representations using dedicated encoders and then feeds these representations into a multimodal encoder, where they are merged and cross-attention is computed.

Mistral-Instruct (Jiang et al. 2023). The Mistral model was trained using the same strategy as the LLAMA model (Touvron et al. 2023), to achieve the best trade-off between performance level and inference budget. Starting with the transformer decoder architecture a series of modifications were made to improve the training stability and efficiency. Additionally, Mistral employs a technique known as *Sliding Window Attention*, which reduces cache memory usage. The model was instruction-tuned on tasks such as *Commonsense Reasoning* and *World Knowledge* between the others. All the training data are publicly available. The model version we used is that with 7 billion parameters, fine-tuned on instruction datasets.

Vicuna (Zheng et al. 2023) The Vicuna model is refined through instruction fine-tuning of a LLAMA 13b model using reinforcement learning with human feedback (RLHF). The fine-tuning data comprises 70,000 conversations shared by users on ShareGPT (Chen et al. 2023). Enhancements include memory optimization and cost reduction techniques and the ability to accurately adhere to instructions in multi-turn dialogues. Evaluation of the chatbot's performance involved identifying eight question categories and collecting ten questions per category. The responses from five chatbots — LLAMA, Alpaca (Taori et al. 2023), ChatGPT (OpenAI 2023), Bard (Google 2023), and Vicuna — were compared, with Vicuna demonstrating superior performance to LLAMA in 90% of cases.

LLAVA-Mistral (Liu et al. 2023). LLAVA is the first vision and language model to employ an instruction-tuning approach. In its foundational version, instruction tuning is accomplished by providing GPT with a prompt consisting of captions and bounding boxes of objects in the image. This prompt aims to generate three types of responses: a conversational exchange, a detailed description, and complex reasoning. Given an input image, the pre-trained visual encoder CLIP extracts its visual features. These features are subsequently converted into language embedding tokens using a linear layer. The language embedding tokens have the same dimensionality as the word embedding space of the language model used, in this case, Mistral. The model training consists of two phases: during the first phase, the weights of both the CLIP model and the textual model are frozen, and solely the weights of the projection layer are fine-tuned. A fine-tuning phase follows, during which the weights of the language model and the linear layer are trained, while those of the CLIP model remain frozen.

LLAVA-Vicuna The LLAVA-Vicuna model has the same architecture as the previous one, differing only in the employment of Vicuna as the language model.

Note that for the 7B and 13B models we used a 4-bit quantized version due to GPU memory constraints.

3.3 Evaluation

In order to evaluate the model's ability to recognize the plausibility of an event we chose to use the perplexity of the models on plausible and implausible sentence pairs and compare the results. The accuracy is then calculated as follows: if the model is

less perplexed by the plausible sentence, the result is considered correct. Otherwise, the result is considered incorrect.

For bidirectional encoder-only models like BERT, VisualBERT, and FLAVA, the pseudo-log-likelihood metric was employed (Salazar et al. 2020). It is computed as the sum of the logarithmic probabilities of each token, considering the context of the sentence. In this case, to mitigate the bias towards multi-token words, we applied an additional mask that covers tokens of the same word located to the right of the target token, as suggested in (Kauf et al. 2022). For decoder-only models, perplexity was simply computed as the average exponential negative log-likelihood. To compare model scores with human judgments expressed on a Likert scale, all values were normalized using a min-max scaler function. Moreover, a plausibility score distribution analysis was conducted to evaluate the similarity between scores for plausible and implausible sentences and the models' ability to distinguish between them. This analysis involves calculating the Pearson correlation coefficient, where negative correlation values indicate good performance. Furthermore, the density of the score distribution is assessed, providing a graphical representation of the degree of overlap between scores for the two classes.

4. Results and Discussion

The models' performance has been evaluated using accuracy, with the results detailed in Table 2 and Table 3.

Table 2
Textual models accuracy on the different datasets

Dataset	Size	Human	BERT	RoBERTa	Mistral-Instr	Vicuna
$DTFit$	395	0.99	0.86	0.89	**0.93**	0.92
$EvAd_{an-in}$	128	1.00	0.93	0.95	**0.99**	0.95
$EvAd_{an-an}$	129	0.95	**0.78**	**0.78**	0.75	0.77
$EvRev$	38	1.00	0.76	0.79	0.89	**0.95**

Table 3
Multimodal models accuracy on the different datasets

Dataset	Size	Human	VisualBERT	FLAVA	LLAVA-Mistral	LLAVA-Vicuna
$DTFit$	395	0.99	0.90	0.86	**0.95**	0.92
$EvAd_{an-in}$	128	1.00	0.93	0.95	**0.97**	0.95
$EvAd_{an-an}$	129	0.95	0.64	0.66	0.74	**0.81**
$EvRev$	38	1.00	0.76	0.79	0.92	**0.95**

Similar to the human judgments, the accuracy levels are higher for all models on the $EventsAdapt_{AN-IN}$ dataset. Specifically, Mistral-Instruct achieves an accuracy of 0.99, while the LLAVA-Mistral model achieves 0.97. The high accuracy levels suggest that when the implausible sentence is impossible, the models can easily identify the concept of event plausibility, thus reaching human-level performance. Moreover, concerning this dataset, we can observe that the VLMs perform comparably with their textual counterparts. On the $EventsAdapt_{AN-AN}$ dataset, where even humans exhibit lower performance *(human=0.95)*, all models perform considerably worse, although above chance. In particular, the VLMs VisualBERT and FLAVA have significantly lower accu-

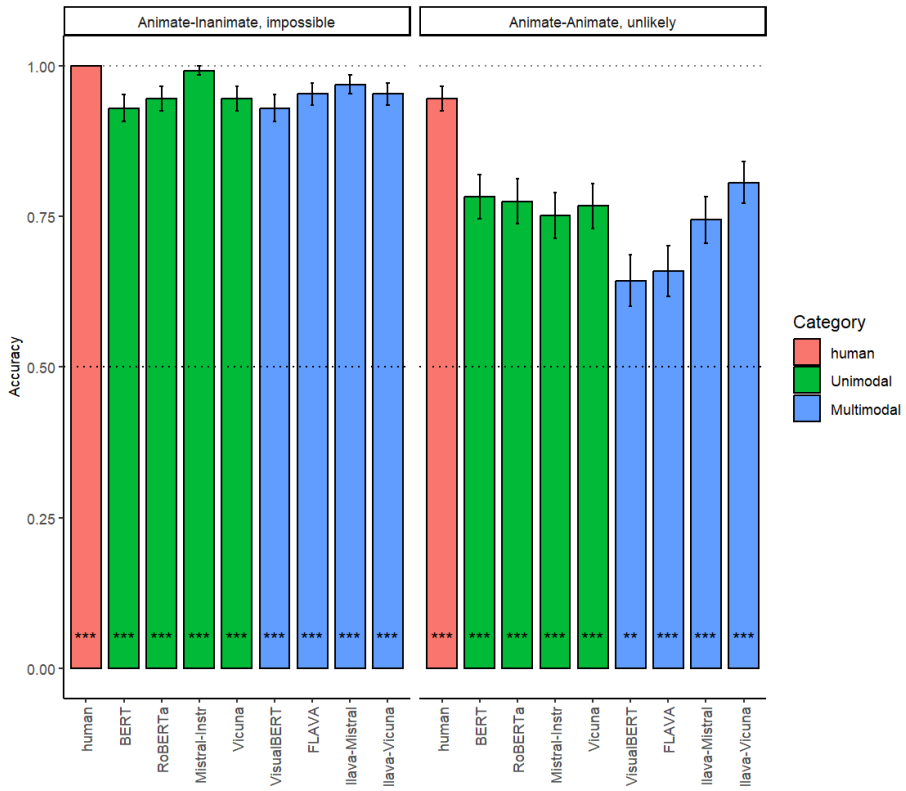

Figure 2
Accuracy plot $EventsAdapt_{AN-IN}$ and $EventsAdapt_{AN-AN}$

racy than the other multimodal models. The others exhibit accuracy levels that are very similar to each other. For readability, we provide a bar chart comparing the accuracy of all models on $EventsAdapt_{AN-IN}$ and $EventsAdapt_{AN-AN}$ in Figure 2.

To gain further insights into the models' behavior, we include an analysis of the data distribution: A density plot of models and human likelihood scores categorized by plausibility (Fig. 3), and a plot showing the correlation of likelihood scores between plausible and implausible sentences (Figs. 4 and 5).

The distribution is similar for both LLMs and VLMs regardless of their size. From the correlation analysis, it's evident that when humans can easily distinguish between the two classes, such as when the implausible sentence violates verb selection preferences (AN-IN), the models assign relatively more distant scores to the two sentences (fig. 4); however, when humans are more uncertain (AN-AN), the models tend to assign very similar scores to the two sentences (Fig. 5).

The density analysis confirms this distribution trend (fig. 3). In AN-IN, there is a clear separation in the distribution for human data and still a relatively clear separation for the models, whereas in AN-AN, the distribution of human scores is less distinct and that of the models shows nearly complete overlap. From the distribution observation, it's apparent that the models struggle to distinguish between plausible and implausible events clearly, and even when assigning a higher likelihood value to the plausible sentence, the difference between them is minimal.

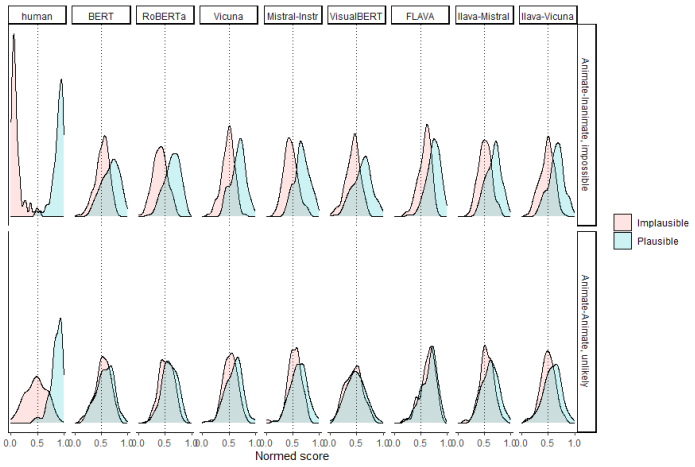

Figure 3
Density plot $EventsAdapt_{AN-IN}$ and $EventsAdapt_{AN-AN}$

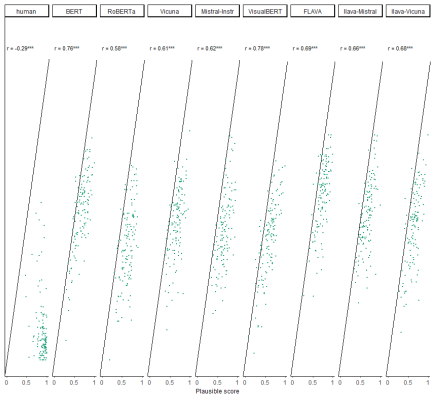

Figure 4
Correlation plot $EventsAdapt_{AN-IN}$

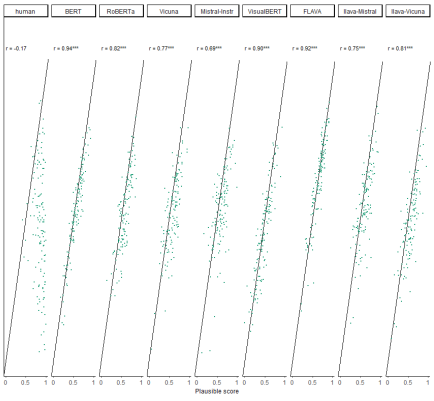

Figure 5
Correlation plot $EventsAdapt_{AN-AN}$

4.1 The impact of model size

Our comparison up to now has not taken into consideration the size of the model, but stark differences exist among the models considered in the experiments. In order to assess the impact of model size, we replicated the experiments on two variants of the same VLM, namely LLaVA-Vicuna. We used the popular 7B and 13B variants. Results are shown in Table 4. No clear trend in favor of the larger model can be identified. We instead observe a variability depending on the dataset.

4.2 The role of concreteness

The level of concreteness of a linguistic element is associated with imageability, defined as the ability of a word to evoke a mental image or sensory experience (Löhr 2024). To define concreteness, we can consider a scale from 1 to 5, where 1 means *"very difficult to*

Table 4
Accuracy comparison between LLAVA-Vicuna7b e LLAVA-Vicuna13b

Dataset	Size	Human	LLAVA-Vicuna7b	LLAVA-Vicuna13b
$DTFit$	395	0.99	0.92	**0.94**
$EventsAdapt_{AN-IN}$	128	1.00	0.95	0.94
$EventsAdapt_{AN-AN}$	129	0.95	0.81	0.80
$EventsRev$	38	1.00	0.95	0.89

imagine" and 5 means *"very easy to imagine"*. Assuming that the degree of concreteness of a sentence is determined by the concreteness of its individual components (Paivio 1971), we divided the sentences into concrete and abstract. First, we annotated the agent, verb, and patient of each sentence with the labels *"concrete"* or *"abstract,"* and then we defined the degree of concreteness of the entire sentence. As a result, sentences expressing abstract events that possessed a high level of imageability were categorized as concrete. For example, the sentence *"The priest celebrated the marriage*" illustrates this point.

Results are shown in Tables 5 and 6, while examples of concrete and abstract sentences for the different datasets are shown in 1. The correlation analysis is instead reported in Appendix 5. With just a few exceptions, VLMs do not show a better ability to deal with concrete events than abstract ones, and textual models address concrete sentences similarly or better than VLMs, *contra* the Dual Coding theory.

Table 5
Textual models accuracy on DTFit and EventsAdapt sentences distinguished by concreteness

Dataset	Size	Human	BERT	RoBERTa	Mistral-Instr	Vicuna
$DTFit\_concr$	350	0.99	0.85	0.90	0.90	**0.92**
$DTFit\_abstract$	45	0.99	0.90	0.87	**0.95**	**0.95**
$EventsAdapt\_concr_{AN-IN}$	97	1	0.95	0.95	**1**	0.95
$EventsAdapt\_abstr_{AN-IN}$	31	1	0.87	0.94	**0.97**	0.94
$EventsAdapt\_concr_{AN-AN}$	65	0.95	**0.82**	0.75	**0.82**	0.75
$EventsAdapt\_abstr_{AN-AN}$	64	0.94	0.75	**0.80**	0.67	0.73

Table 6
Multimodal models accuracy on DTFit and EventsAdapt sentences distinguished by concreteness

Dataset	Size	Human	VisualBERT	FLAVA	LLAVA-Mistral	LLAVA-Vicuna
$DTFit\_concr$	350	0.99	**0.90**	0.86	**0.94**	0.92
$DTFit\_abstract$	45	0.99	0.92	0.92	**0.97**	**0.97**
$EventsAdapt\_concr_{AN-IN}$	97	1	0.94	0.95	**0.98**	0.95
$EventsAdapt\_abstr_{AN-IN}$	31	1	0.90	**0.97**	0.94	**0.97**
$EventsAdapt\_concr_{AN-AN}$	65	0.96	0.70	0.67	**0.82**	**0.82**
$EventsAdapt\_abstr_{AN-AN}$	64	0.94	0.56	0.62	0.66	**0.78**

4.3 The role of visual inputs

In order to evaluate the performances of VLMs on multi-modal stimuli (i.e., images + texts), we use the EventsRev dataset. Recall that each sentence is associated with an image depicting what is described in the sentence. All images are black-and-white sketches with the same style. Previous works indicated that encoder-only VLMs generally underperformed on the specific dataset, and did not highlight specific trends in performances with or without including the images as inputs (Cassese, Bondielli, and Lenci 2023). Accuracy values were generally lower than those obtained on other datasets, and the inclusion of images affected them only for one of the tested models. Here we include additional decoder-only models in the experimentation. We start from the following assumption. If the model is properly encoding the information contained in the image and the text, we would expect scores for plausible sentences to remain rather similar with or without the image, and scores for implausible ones to change as grounding the sentence with the related image would make the situation described in the following tokens less implausible. This would be reflected, in our proxy-task, in higher accuracies on the text-only task, and significantly lower values for the text+image task, as both sentences have more similar PPL/PLL values. Recall in fact how the accuracy is computed: we consider a sentence pair (S_p, S_i) to be correctly modeled if $PLL(S_i) < PLL(S_p)$ (or $PPL(S_i) < PPL(S_p)$), that is the model consider the plausible sentence as more likely than the implausible one.

As shown in Table 7, the accuracy values for the text-only task are significantly lower in encoder-only models than in decoder-only ones. As for the text+image task, the performances either remained unchanged (for the FLAVA model) or were lower than if using text alone. In the case of LLAVA-Vicuna, the decrease in performance is quite significant, with a reduction of up to 16 percentage points. To further examine this issue, we consider how the distribution of PPL/PLL values changes by including the images in the inputs. Results for LLAVA-Vicuna are shown in Figures 6 and 7. Other models are reported in Appendix 5. We see that including images in the input slightly decreases the difference between mean values and consistently increases in-group standard deviation, thus making the two distributions more similar. We show this in the example in Figure 8, in which we report one of the stimuli and the PPL scores for LLAVA-Mistral with and without including the corresponding picture in the input.

Table 7
Accuracy of VLMs on EventsRev with $(t + i)$ and without (t) images in the input.

Dataset	**VisualBERT**	**FLAVA**	**LLAVA-Mistral**	**LLAVA-Vicuna**
$EventsRev_t$	0.76	0.79	0.92	**0.95**
$EventsRev_{t+i}$	0.61	0.79	**0.84**	0.79

Hypotheses for these outcomes are further discussed in Section 4.5.

4.4 Error Analysis

To determine if there are differences in the types of errors made by different models and the extent to which these errors are due to violations of plausibility, isolated analyses were performed. These analyses considered a) only the sentences that all models got

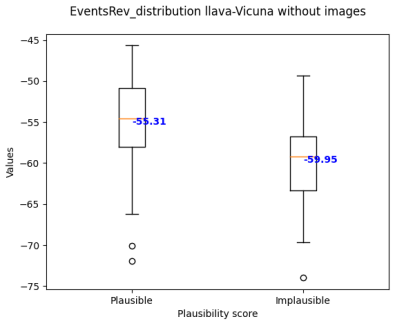

Figure 6
PPL values distribution for plausible and implausible sentences using LLAVA-Vicuna without images in the input.

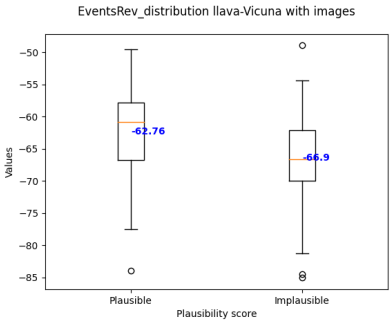

Figure 7
PPL values distribution for plausible and implausible sentences using LLAVA-Vicuna with images in the input.

		PPL	
Image	**Text**	**Text Only**	**Text + Image**
	The pirate is kidnapping the princess	-44.72	-72.5
	The princess is kidnapping the pirate	-51.16	-71.9

Figure 8
Comparison of PPL scores for a sentence pair with and without including the corresponding image as input.

wrong; b) those that only the multimodal model got wrong; and c) those that only the unimodal models got wrong. The analyses were repeated across all datasets.

4.4.1 EventsAdapt

On the EventsAdapt dataset, all models achieve very high performance for the Animate-Inanimate subset, resulting in very few errors. In this case, it is observed that the pairs of similar unimodal and multimodal models (BERT and VisualBERT, LLAVA-Mistral and Mistral, LLAVA-Vicuna and Vicuna) tend to make errors in the same sentences. However, the multimodal models make a few more mistakes compared to their unimodal counterparts.

In the accuracy analyses, it was observed that the human accuracy rate for the Animate-Animate subset of EventsAdapt is 0.95. This indicates that there are sentence pairs where the atypical sentence is assigned a higher score than the typical one, leading to a negative difference in human scores between the two sentences. Models also tend to make mistakes in these sentences, as exemplified by the following pair: *"The reviewer criticized the right-winger"*, *"The right-winger criticized the reviewer"*, for which the human score difference was -0.04. Besides these, there are also sentences where the difference

in the average score assigned by humans is positive, but very low (≤ 0.10) (*"The genius shocked the cousin."*, *"The cousin shocked the genius"*). For these instances, we observe two key characteristics: first, determining which scenario is more likely is challenging without context; second, these situations are not prototypical, making it difficult to discern which scenario is more plausible.

The cases where humans exhibit high confidence in assigning different scores to the two sentences, while the models make errors, are particularly interesting. In these cases, we observe that the more models share similar architectures and components (for example, the same language model), the more similar their performance tends to be. In general, models tend to make more mistakes with sentences that denote less prototypical events, have low imaginability, and are more abstract. For example, all multimodal models fail to categorize the sentence *"the environmentalist warned the tobacconist"*. In this case, the association between *"environmentalist"*and *"tobacconist"* is weak, whereas it would have been more typical if we had *"smoker"* instead of *"tobacconist"*. Additionally, the verb *"to warn"* makes the sentence abstract and more difficult to imagine. However, a clear difference in behavior between unimodal and multimodal models cannot be established. Indeed, the sets of sentences incorrectly predicted solely by unimodal models and those solely by multimodal models are empty. In contrast, the error distributions for the models BERT, RoBERTa, VisualBERT, and FLAVA overlap, as do those for the models Vicuna, Mistral, llava-Vicuna, and llava-Mistral. Consequently, the error distribution of the models on this specific task depends more on the language model architecture and the training procedure used than on the interaction between the unimodal and multimodal components.

4.4.2 DTFit

In the DTFit dataset, a common error is failing to recognize the typical sentence among two plausible variants, as seen with the pair *"The guest held the drink"* (typical) and *"The guest held the camera"* (atypical). In such cases, the strength of the verb-patient association is crucial: for instance, when considering the verb *"to sign"*, the phrase *"to sign the agreement"* is more intuitively associated than *"to sign the paint"*. This could cause the model to choose the implausible sentence *"The painter signed the agreement"* instead of the plausible sentence *"The painter signed the paint"*. No specific errors related to a particular category of models were identified.

4.4.3 Error distinguished by concreteness

To accurately interpret the results obtained from concrete and abstract sentences, it is important to consider two factors: the ratio of concrete to abstract sentences across different datasets and the method used to categorize them.

In the DTFit dataset, the number of abstract sentences is significantly lower than that of concrete sentences, with a ratio of approximately 13 abstract sentences for every 100 concrete ones. Consequently, the fact that all models except RoBERTa perform better on abstract sentences may be influenced by the disparity in dataset sizes. The same occurs in the EventsAdapt Animate-Inanimate dataset, although to a lesser extent. It is more interesting to observe the results in the EventsAdapt Animate-Animate dataset, where the distribution of concrete and abstract sentences is nearly equal (65 concrete sentences and 64 abstract sentences). In this dataset, all models tend to achieve higher accuracy on concrete sentences (tables 5, 6).

If we examine the sentences on which the models make errors, it becomes apparent that the abstract sentences in the dataset often represent less prototypical events.

Therefore, understanding which variant is more likely requires reasoning and abstraction abilities. For example, the sentence *"The arsonist alarmed the vendor"* is difficult to interpret without context. When we read it, we might imagine various scenarios, such as a threat, an ongoing fire, or behavior that alarms the vendor. In contrast, a concrete sentence like *"The zookeeper fed the giraffe"* is much easier to interpret.

Ultimately, it is observed that the disparity in dataset sizes impacts model accuracy, obscuring the inherent difficulty models face with abstract sentences.

4.4.4 EventsRev: Error analysis in multimodal models with and without images

The EventsRev dataset contains only concrete sentences that describe prototypical events. Consequently, there are no specific patterns in the types of errors made by the models. However, some observations can be made. Let's examine models with similar architectures to identify any recurring patterns. Notably, the BERT and the VisualBERT models with only textual input almost always make the same mistakes in the same sentences. When visual input is added to VisualBERT, some errors remain common with the previous results, but new errors emerge.

Regarding large models, it is observed that the Mistral and Vicuna models, as well as the LLAVA-Mistral and LLAVA-Vicuna with only textual input, rarely make errors. However, when visual input is added to the multimodal models, the errors double in the LLAVA-Mistral model and quadruple in the LLAVA-Vicuna model (table 7).

In conclusion, we can say that all the models tend to increase the number of errors when visual input is added.

4.5 Discussion

The results of the experiments provide us with some interesting insights, especially in light of what was already found in (Cassese, Bondielli, and Lenci 2023).

First, we observe a significant difference in the behavior of encoder-only and decoder-only models. While for VLMs based on encoder-only LLMs, the performances are consistently inferior to the text-only counterparts, the same cannot be said for more modern decoder-only based models. In fact, we observe that in the latter case, VLMs performances are on par or better than those of the LLM counterpart on the text-only task. This is also independent of the LLM performances: we do not observe a linear relationship between LLM performances and its VLM version. This finding compared with previous results (Cassese, Bondielli, and Lenci 2023), suggests that more modern models exhibit a different behavior as they are better equipped to solve the task. This may be an indication that these models behave less as bag-of-words models if compared to older, encoder-only based models. This may be because a current trend in the literature consists of using a frozen image encoder and a strong pre-trained LLM as the basis for generative VLMs. Decoupling the training of text and image encoders during training may thus partly prevent the LLM from being influenced to a bag-of-words behavior as seen for encoder-only models (Cassese, Bondielli, and Lenci 2023). On the other hand, it is worth emphasizing that VLMs never significantly outperform textual LLMs, suggesting the limited added value provided by visual information over linguistic ones in current models, at least in the task of event plausibility recognition.

Second, we observe that contrary to current trends in the literature, using larger models does not appear to significantly change or consistently improve performances. The 7B and 13B variants of LLAVA-Vicuna perform very similarly, with a difference of a few points in favor of one or the other depending on the dataset 1. It would be

interesting to further test this hypothesis on LLMs with bigger orders of magnitude of parameters (e.g., 70-80B to 100-200B models).

Third, we observe how VLMs exhibit lower performance on the more challenging sentences in the $EventsAdapt_{AN-AN}$ dataset, displaying reduced accuracy and a high correlation between plausible and implausible sentences. This is consistent with previous findings (Kauf et al. 2022) and suggests that even the most recent models are still not able to approximate human-level performances for all degrees of complexity of the input.

Finally, we find some significant changes in performances when including also the corresponding visual stimulus to the input. More specifically, performances either remain the same (for FLAVA) or drastically decrease, especially for the otherwise best-performing model, namely LLAVA-Vicuna. This finding is especially interesting as it may be interpreted from two opposing perspectives. On the one hand, we could posit that this suggests that VLMs still struggle to identify relationships between elements and understand the action in an image. This trend would be further confirmed by the high correlation between PLL/PPL scores for pairs of (S_p, S_i) in VLMs, particularly noticeable for sentences where the implausible sentence is generated by reversing the order of the agent and patient in the sentence. This could be taken as a suggestion that not only encoder-only models but also large decoder-only models like LLAVA, continue to exhibit the same limitation in interpreting compositional information. We could also argue that an unlikely event is not more likely if it is grounded by an image. A human would probably consider the event of a princess kidnapping a pirate much more unlikely than the opposite, even after directly witnessing it. The observed shift in PPL values towards higher ones for implausible texts + images pairs would suggest that the model does not reason in the same way, and thus is less effective at characterizing the plausibility of an event.

On the other hand, however, this finding may be indicative of the positive impact of including the images for the model's global understanding in decoder-only models. As mentioned in Section 4.3, we could hypothesize that more similar values of PPL/PLL for a sentence pair may be precisely due to the images providing crucial context and grounding to the model: implausible sentences could be considered as surprising for the model as plausible ones by the model once it has been given the context in the form of the image, actually depicting the implausible event; conversely, the same decrease in PPL/PLL may not be observed for plausible sentences, for which the image may not be as crucial for interpreting the text, as it is generally likely to begin with. Thus, the PPL/PLL of implausible sentences and images may be slightly higher in some cases than the respective plausible ones, lowering the accuracy. The box plot results presented in 4.3 may be considered as preliminary evidence in favor of this hypothesis. This, in addition to the better overall performances, would indicate that newer decoder-only models are better able to model the contents of an image and relate it to its textual context, and vice versa.

This may be an interesting finding that delves deeper into the problem of modeling multimodal aspects in VLMs and leaves the door open for further investigation. However, we must point out that a relevant limitation of this analysis that prevents us from deriving stronger conclusions is posed by the small sample size of the EventsRev dataset, which includes only 38 samples. In order to better address this aspect, it is crucial to further test both hypotheses by employing larger samples of data and different metrics. The limited number of examples available for this analysis represents its main limitation. Therefore, it will be necessary to construct a larger dataset to conduct more in-depth analyses and achieve more robust and generalizable results.

5. Conclusions

In this paper, we assessed language and vision-language models on identifying event plausibility, including several dimensions of analysis in terms of models' characteristics. We evaluated the models on three different datasets of sentence pairs on the task of deciding whether the plausible sentence in the pair is more likely than the implausible one, using the perplexity of the model as a proxy metric. We experimented with different model architectures and different model sizes. We provided an in-depth comparison that takes into account three aspects: first, the concreteness of the stimuli; second, whether the implausible stimulus in the pair actually violates verb selection preferences or not (e.g., *"a stone is arresting a thief"* vs *"a thief is arresting a cop"*); third, we evaluated the impact of including images corresponding to the stimulus in the input to VLMs.

Our findings are partly in line with previous works (Cassese, Bondielli, and Lenci 2023), but provide further indications on the behavior of VLMs and highlight the increased effectiveness of state-of-the-art VLMs in properly modeling linguistic knowledge. First, we see that newer models, generally based on a decoder-only architecture, outperform encoder-only ones. Second, increasing the model size does not consistently improve its performance. Third, we notice that contrary to (Cassese, Bondielli, and Lenci 2023), state-of-the-art VLMs appear to be on par or slightly better than their LLM counterparts, suggesting that the bag-of-words behavior exhibited by encoder-only VLMs is less noticeable. Fourth, we observe a significant drop in performances, in terms of selection accuracy of the more plausible sentence over the implausible one, when including images in VLMs inputs. We hypothesize that this may be due to better modeling of texts and images together: when the model is provided with a context (i.e., the image depicting the stimulus), it is less "surprised" by the implausible stimulus with respect to considering text only, thus leading to a decrease in accuracy.

In future works, we intend to further investigate this latter finding, as it may shed more light on the capabilities of decoder-only VLMs. We intend to do so by i.) considering larger datasets that allow us to provide stronger evidence of the hypothesis, ii.) using alternative metrics to perplexity, and iii.) evaluating the extent to which providing non-matching images affects the performances.

Acknowledgments

Research partially supported by the Italian Ministry of University and Research (MUR) in the framework of the PON 2014-2021 "Research and Innovation" resources – Innovation Action - DM MUR 1062/2021 - Title of the Research: "Modelli semantici multimodali per l'industria 4.0 e le digital humanities.", and by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU program. The work of Maria Cassese has been carried out in the frame of the ITSERR project, financed by the European Union under the NextGenerationEU funding scheme.

Appendix A: Correlation of likelihood between plausible and implausible sentences considering concreteness

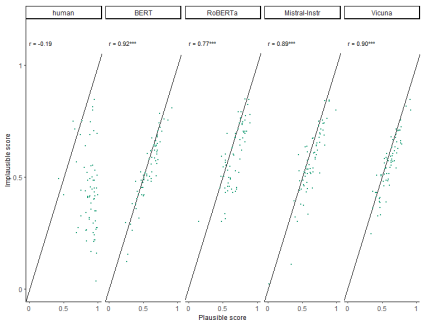

Figure 1
$EventsAdapt^{concr}_{AN-AN}$ unimodal models

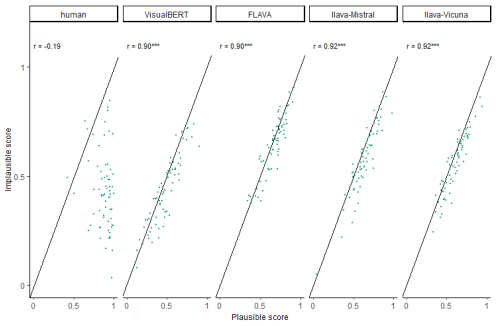

Figure 2
$EventsAdapt^{concr}_{AN-AN}$ multimodal models

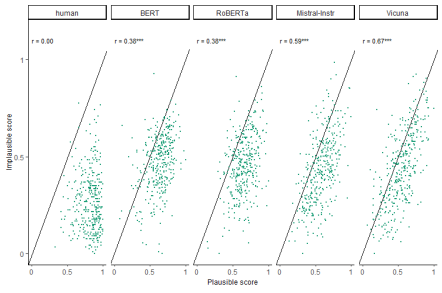

Figure 3
$DTFit^{concr}$ unimodal models

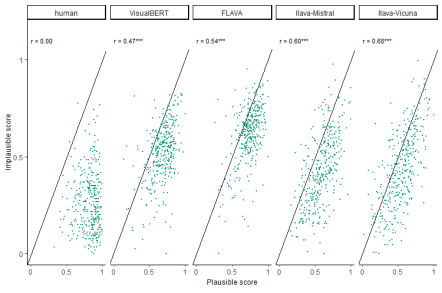

Figure 4
$DTFit^{concr}$ multimodal models

Appendix B: Perplexity changes with image inputs

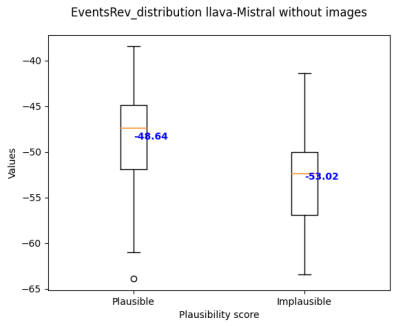

Figure 1
PPL values distribution for plausible and implausible sentences using LLaVa-Mistral-7B without images in the input.

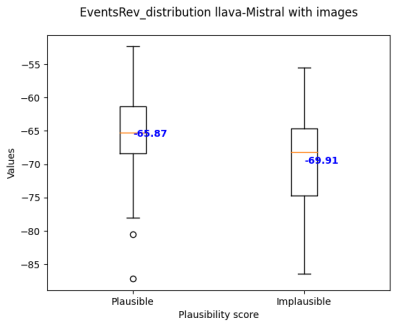

Figure 2
PPL values distribution for plausible and implausible sentences using LLaVa-Mistral-7B with images in the input.

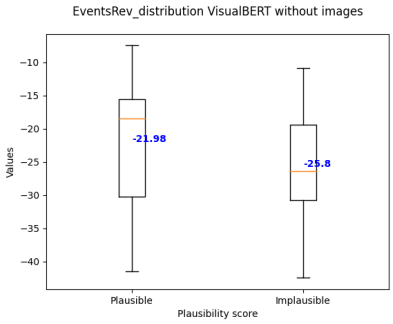

Figure 3
PPL values distribution for plausible and implausible sentences using VisualBERT without images in the input.

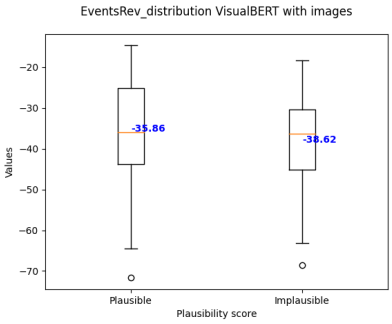

Figure 4
PPL values distribution for plausible and implausible sentences using VisualBERT with images in the input.

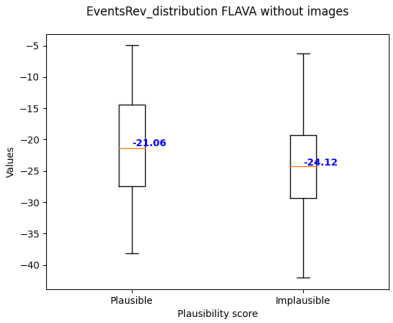

Figure 5
PPL values distribution for plausible and implausible sentences using FLAVA without images in the input.

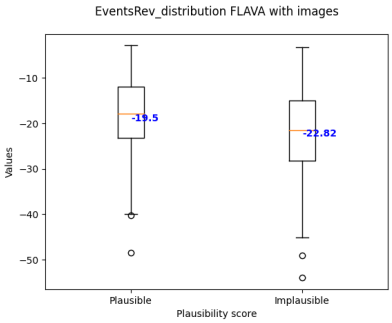

Figure 6
PPL values distribution for plausible and implausible sentences using FLAVA with images in the input.

References

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, page 136–145, Jeju Island, Korea, July.

Cassese, Maria, Alessandro Bondielli, and Alessandro Lenci. 2023. Assessing language and vision-language models on event plausibility. In *Proceedings of the 9th Italian Conference on Computational Linguistics*, Venice, Italy, November 30 - December 2.

Chen, Lin, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR 2021 - The Ninth International Conference on Learning Representations*, Online, May.

Fedorenko, Evelina, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*,

203:104348.
Feng, Yansong and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99, Los Angeles, California, June. Association for Computational Linguistics.
Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November.
Google. 2023. Bard, "https://bard.google.com/".
Gordon, Jonathan and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, San Francisco, California, USA, October-November. Association for Computing Machinery.
Harnad, Stevan. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.
Ivanova, Anna A., Zachary Mineroff, Vitor Zimmerer, Nancy Kanwisher, Rosemary Varley, and Evelina Fedorenko. 2021. The Language Network Is Recruited but Not Required for Nonverbal Event Semantics. *Neurobiology of Language*, 2(2):176–201, 03.
Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
Kauf, Carina, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan S. She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2022. Event knowledge in large language models: the gap between the impossible and the unlikely. *arXiv preprint arXiv:2212.01488*.
LeCun, Yann, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
Löhr, Guido. 2024. Does the mind care about whether a word is abstract or concrete? why concreteness is probably not a natural kind. *Mind & Language*, 39:627–646.
Marconi, Diego. 1997. *Lexical Competence*. The MIT Press, Cambridge, MA.
Matsuki, Kazunaga, Tracy Chow, Mary Hare, Jeffrey L. Elman, Christoph Scheepers, and Ken McRae. 2011. Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4):913.
McRae, Ken and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, 3(6):1417–1429.
Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.
Morris, Charles William. 1938. Foundations of the theory of signs. In *International encyclopedia of unified science*. Chicago University Press, pages 1–59.
OpenAI. 2023. Chatgpt. `https://openai.com/blog/chatgpt/`.
Paivio, Allan. 1971. Imagery and language. In *Imagery*. Elsevier, pages 7–32.
Pavlick, Ellie. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.
Pedinotti, Paolo, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the cat drink the coffee? challenging transformers with generalized event knowledge. *arXiv preprint arXiv:2107.10922*.
Pezzelle, Sandro, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.
Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July. Association for Computational Linguistics.

Shekhar, Ravi, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada, July. Association for Computational Linguistics.

Sherstinsky, Alex. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Singh, Amanpreet, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. Flava: A foundational language and vision alignment model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629, June.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, , and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Thrush, Tristan, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *arXiv preprint arXiv:2204.03162*.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vassallo, Paolo, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2018. Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality. In *LREC 2018 Workshop on Linguistic and Neurocognitive Resources (LiNCR)*, Miyazaki, Japan, May.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Long Beach, California, USA, December. Curran Associates Inc.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Yuksekgonul, Mert, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Zellers, Rowan, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, Long Beach, CA, USA, June.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.