Contextualising (Im)plausible Events Triggers Figurative Language

Annerose Eichel, Tonmoy Rakshit, Sabine Schulte im Walde
Institute for Natural Language Processing
University of Stuttgart
{annerose.eichel,tonmoy.rakshit,schulte}@ims.uni-stuttgart.de

Abstract

This work explores the connection between (non-)literalness and plausibility at the example of subject-verb-object events in English. We design a systematic setup of plausible and implausible event triples in combination with abstract and concrete constituent categories. Our analysis of human and LLM-generated judgments and example contexts reveals substantial differences between assessments of plausibility. While humans excel at nuanced detection and contextualization of (non-)literal vs. implausible events, LLM results reveal only shallow contextualization patterns with a bias to trade implausibility for non-literal, plausible interpretations.

Keywords: plausibility, (non-)literalness, human vs. LLM generation

1. Introduction and Related Work

How plausible do you judge a situation where the *heat catches a cyclist*? Based on the literal reading of this described situation, one might expect that the majority of human annotators would consider this situation as rather implausible. Yet, previous research has demonstrated that humans are eager to interpret such a seemingly implausible event. In the current study, we investigate to which extent the interpretation of implausible events is driven by the potential to frame them in a non-literal context. For instance, the above situation has been contextualized in figurative example sentences such as *During the intense desert race, the scorching heat caught the cyclist off guard, forcing him to stop for water and shade.* and *The unexpected heat caught the cyclist unaware.*

As the example illustrates, interpreting and distinguishing plausible from implausible events is a crucial and non-trivial building block of natural language. A range of work has explored semantic plausibility using subject-verb-object (svo) events in English leveraging embedding-based neural networks (Wang et al., 2018), transformers (Porada et al., 2019; Emami et al., 2021; Porada et al., 2021), and LLM-based methods (Kauf et al., 2024). However, prior research so far explicitly focused on literal events (Wang et al., 2018) or other aspects of plausibility. Our work addresses this gap and explores the connection between figurative language and plausibility. To do so, we adopt definitions from previous work (Wilks, 1975; Resnik, 1993; Wang et al., 2018) and consider plausibility in a binary setting. *Plausible* events include not only highly typical events but also untypical events (Wilks, 1975), potentially novel events (Wang et al., 2018), and seemingly trivial events such as "a person breathes" that are not necessarily attested in an existing corpus (Gordon and Van Durme, 2013). In comparison, fully implausible events do not allow any semantically valid interpretation; neither a literal nor a figurative reading, for example, through creative metaphors (Griciūtė et al., 2022).

For our study, we rely on a small subset of svo event triples that were previously annotated as (im)plausible (Eichel and Schulte im Walde, 2023). Crucially, the original triples are balanced with regard to the degree of concreteness vs. abstractness of the involved constituent words, because concepts can be described in accordance with the way people perceive them (Barsalou and Wiemer-Hastings, 2005; Brysbaert et al., 2014): Concrete concepts such as *trampoline* can be seen, heard, touched, smelled, or tasted. In contrast, abstract concepts such as *realism* cannot be perceived with the five senses. In between these two extremes on the scale, mid-range concepts such as *punctuality* are situated. The provided abstractness information allows us to connect not only (im)plausibility to (non)literalness but to additionally integrate an interaction with conceptual abstractness, thus implicitly relating to Conceptual Metaphor Theory (Lakoff and Johnson, 1980) as a mapping from abstract to concrete concepts to trigger metaphorical meanings as a special case of non-literal language.

More specifically, we make use of 411 svo triples with plausibility judgments, and ask humans and LLMs to make a binary judgement about their figurative language, plus providing example sentences. Our novel dataset contains a total of 6,497/14,555 judgments and 6,497/3,288 unique sentences generated by humans/LLMs.[1] We use the collected judgements and sentences to compare human and LLM generations. In the context of plausibility, humans have been observed to tend towards sense-making with great nuance and willingness to interpret even whimsical sentences (Griciūtė et al.,

[1] www.github.com/AnneroseEichel/NLE2026

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 93–105
11 May 2026.

2022; Eichel and Schulte im Walde, 2023). Regarding LLMs, while they are equipped for semantic interpretation with (world) knowledge learned through distributional patterns in vast amounts of training data, almost all of their data are plausible. We thus formulate the following research questions:

- RQ1: Does figurative language interact with event plausibility, and how does this interaction relate to the abstractness of the event constituents?
- RQ2: How do human annotations compare to LLM judgments regarding figurative language and event plausibility?
- RQ3: Which qualitative differences can be observed for human vs. model-produced contextualizations of (non-)literal implausible events?

Our contributions are threefold: (i) We present a collection of judgments and example contexts from humans and four LLMs, using a systematic setup of plausible and implausible event triples in combination with abstract and concrete constituent categories. (ii) We provide detailed insights into human vs. LLM judgments for predicting (non-)literalness, and (iii) we conduct a careful analysis of human- and model-generated contexts as well as repair mechanisms for seemingly implausible events.

2. Data

We use the plausibility dataset PAP (Eichel and Schulte im Walde, 2023). PAP encompasses a balanced set of 1,733 subject-verb-object triples in English extracted from Wikipedia (originally plausible) and automatically perturbed triples (originally implausible). All events are labeled by each component's concreteness ranging from abstract (a), over mid-range (m), to concrete (c) (Brysbaert et al., 2014). PAP is balanced across all possible combinations of abstractness such as events consisting of only highly concrete words such as "*person calls town*" (ccc) or fully mixed events such as "*career reestablishes chicken*" (amc). Triples are annotated through crowd-sourcing with subjective assessments of plausibility on a degree scale (1–5) ranging from implausible to plausible. PAP ratings include raw annotations as well as original plausibility labels, and provide clear majority-based ($\geq 70\%$) aggregations. For this study, we use a subset satisfying the following criteria: across all abstractness combinations, we draw a random sample of event triples where (i) original and human-annotated majority label correspond to each other such as "*album breaks genre*" (orig.: *plausible*; PAP maj.: *plausible*), and (ii) original and human-annotated majority label differ such as "*collection needs autonomy*" (orig.: *implausible*; PAP maj.: *plausible*). An overview is shown in App. A.1, Table 4.

3. Methods

For each svo event in our dataset sample, we collect judgments and example sentences from both humans and LLMs. We ask humans to (1) select a label for whether an event is figurative, literal, or neither, based on the combination of component meanings, and (2) produce an example contextualizing the event (only if figurative or literal), or a sentence contextualizing an altered event if neither (cf. App. A.2, Figure 3 for annotation instructions). Then, we prompt LLMs to complete the same tasks.

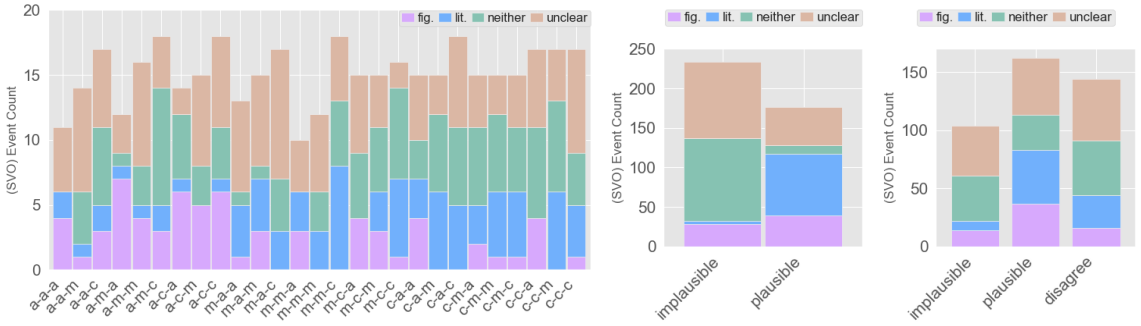
Human Annotation We use Prolific and Google Forms as study tools. Participants are required to reside in the UK, US, Ireland, or Australia and hold corresponding citizenship, speak English as their primary language, and have a Prolific approval rate of $\geq 98\%$. Items are shown in batches of ≈25 items with one target shown per page. For each item, we collect $16^{\pm 1.45}$ responses from a total of 240 annotators. We make sure that each annotator contributes $< 1\%$. to the collection. Our annotator sample has a median age of 38 years, is slightly skewed towards female (56.7%) over male annotators (43.3%), and resides mainly in the US (57%). For full demographic details, cf. App. A.2. Across abstractness combinations, we obtain a total of 6,497 judgments and example sentences.

Modeling For model predictions, we use the Instruct versions of four LLMs. We focus on moderate parameter count and test four multilingual model families: Qwen3-4B (Qwen Team, 2025), Gemma3-4B (Gemma Team, 2025), Mistral-7B (Jiang et al., 2023), and Llama3.1-8B (Grattafiori et al., 2024). The LLM prompts for label and text generation are based on instructions for humans with (few-shot) and without (zero-shot) examples (cf. App. A.3, Figure 6 and Figure 5 for prompts). (1) For label prediction, we aggregate results across five model runs with different random seeds and three prompts with varying phrasing and output formats. (2) For example sentence generation, we replicate human instructions as closely as possible and obtain generations for one seed. All prompting is performed with model default settings and a 64 token limit.

4. Results

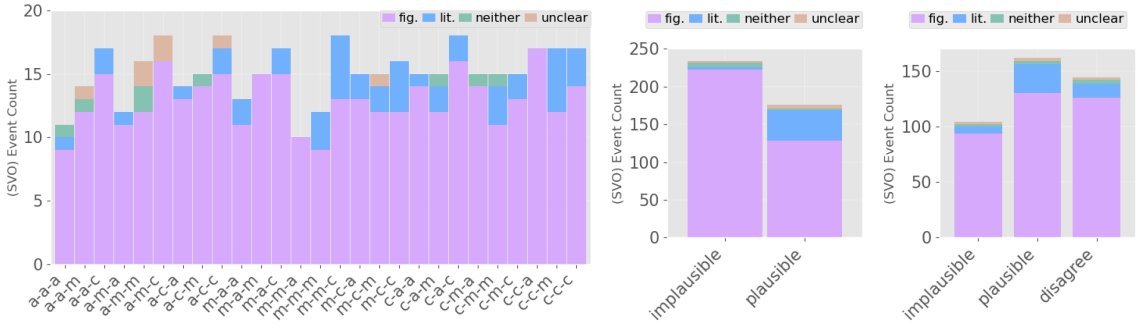
4.1. Figuratively, literally, unclear, or actually not plausible at all?

To explore RQ1, i.e., how figurative language interacts with event plausibility and event constituent abstractness, we first provide a deep-dive into human judgments. To shed light on RQ2, we then assess how human annotations compare to SOTA LLMs' judgments.



(a) **Human** judgements across abstractness comb. (b) orig. label (c) PAP rating

Figure 1: Analysis of **human** figurative majority-assigned labels assigned across (a) 27 **abstractness combinations** ranging from most abstract on the left to most concrete one the right, (b) in comparison with **original labels** used to create PAP, and (c) **PAP majority ratings**.



(a) `Qwen3-4B` judgements across abstractness comb. (b) orig. label (c) PAP rating

Figure 2: Analysis of `Qwen3-4B` figurative labels (zero-shot), similarly to Figure 1.

Human Judgments To investigate the relationship between **abstractness** as well as original and PAP **plausibility** ratings and figurative meanings of the targets, we visualize the number of svo events judged by the majority of participants. Here, majority is defined as $\geq 60\%$. Whenever no majority is reached, we assign the label *unclear*. Figure 1 presents three perspectives on the distribution of judgments. Across plots, label categories are: *figurative* (violet), *literal* (blue), *neither* (no contextualization is possible, green), and *unclear* (orange).

We first look at the interplay between the **abstractness of the svo events, and their perception as figurative vs. literal language**. Across the x-axis, we observe a clear trend. More concrete events (e.g., (*ccc*), (*cam*), (*cac*)) are judged more literally, while more abstract events (e.g., (*aaa*), (*ama*), (*mca*)) are judged more figuratively. More specifically, subject and object abstractness exert greater influence on (non-)literalness. In particular, concrete or mid-scale verbs in such events lead to predominantly literal readings of events, confirming prior work (Khaliq et al., 2024; Knupleš et al., 2026). In turn, we observe a limited influence of verb abstractness. Items for which neither a figurative nor a literal reading or a context could be inferred such as "*payload lives blowout*" or "*city folds fruit*" accumulate on the more concrete end of the scale. Here, concrete objects strongly influence events that are otherwise abstract.

Next, we focus on the relationship between **plausibility and figurative language**. We consider both the original labels underlying the PAP dataset, and the majority PAP ratings. Originally plausible items such as "advance guarantees freedom" are judged more figuratively than originally implausible items such as "*copy inflates disbelief*". While the difference is rather small for original labels, a larger disparity is observed for PAP ratings. When inspecting targets judged literally such as "*owner secures trademark*", a virtual clear-cut between plausible and implausible items emerges. While both original and majority PAP plausible items are judged mainly literally, implausible items are virtually never perceived as literal.

Following our qualitative inspection, we further **quantify whether assigned (non-)literal labels are related with event abstractness, original label, or PAP ratings** using χ^2 tests of independence (Pearson, 1991). Strength of association is examined with Cramér's V (Cramér, 1999). An overview of results is presented in Table 1. We find significant associations ($p < .001$) between abstractness combinations and figurative/literal labels, with a mod-

		Human		Qwen3-4B		Llama3.1-8B		Mistral-7B		Gemma3-4B	
Figurative	df	χ^2	V	χ^2	V	χ^2	V	χ^2	V	χ^2	V
ABSTRACTNESS	26	**63.64*** **	0.39	23.10	0.24	48.13**	0.34	41.50*	0.32	42.82*	0.32
ORIG. LABEL	1	6.91**	0.13	37.68	0.30	0.17	0.02	4.64*	0.11	33.10	0.28
PAP RATING	3	13.49**	0.18	6.20	0.12	2.88	0.08	1.59	0.06	7.68	0.14
Literal											
ABSTRACTNESS	26	43.93**	0.33	34.81	0.29	50.88**	0.35	42.67*	0.32	43.75*	0.33
ORIG. LABEL	1	**111.33*** **	0.52	45.72	0.33	0.46	0.03	4.66*	0.11	32.23	0.28
PAP RATING	3	**17.29*** **	0.21	7.85*	0.14	2.25	0.07	5.42	0.11	8.09*	0.14
Neither											
ABSTRACTNESS	26	32.80	0.28	27.33	0.26	23.33	0.24	-	-	32.86	0.28
ORIG. LABEL	1	71.96	0.42	0.45	0.03	1.53	0.06	-	-	1.19	0.05
PAP RATING	3	13.73**	0.18	0.04	0.01	0.48	0.03	-	-	2.16	0.07
Unclear											
ABSTRACTNESS	26	29.59	0.27	32.56	0.28	28.31	0.26	39.61*	0.31	40.18*	0.31
ORIG. LABEL	1	8.23**	0.14	1.33	0.06	0.80	0.04	0.32	0.03	1.50	0.06
PAP RATING	3	4.17	0.10	0.15	0.02	2.10	0.07	0.59	0.04	5.42	0.11

Table 1: Associations between figurative language and abstractness, original label, or PAP ratings. χ^2 indicates *significance* ($p < .05$:**, $p < .01$:**, $p < .001$: ***) and Cramér's V measures *strength* of association. Model results are based on zero-shot prompts.

erate effect size. This finding further underlines previous work (Khaliq et al., 2024; Knupleš et al., 2026) focusing on verb-object (v,o) pairs where they find an increase in figurative majority-based judgments predominantly influenced by object abstractness. We further find a significant association ($p < .001$) between original and literal labels, with a strong effect size. For literal labels and majority PAP, we also observe a statistically reliable association ($p < .001$) albeit with a weaker effect size.

Human vs. LLM Judgments We compare majority-assigned label distributions obtained by humans vs. four SOTA LLMs where majority is defined as $\geq 60\%$. Results across models and prompt types are shown in Table 2 with performance equally low for all four models, i.e., they mostly disagree with human judgments. When visualizing majority-assigned labels, we observe clear differences across models and prompt types. For reasons of space, we illustrate model results at the example of Qwen3-4B results in Figure 2 and provide a full overview in App. B, Figure 7.

Zero-shot prompts trigger Gemma, Qwen, and Mistral to assign overwhelming majorities of plausible, and specifically figurative readings. This holds for both originally plausible and implausible events. A notable exception is Llama which assigns significantly more literal interpretations across original labels and abstractness combinations. This trend is only partially observable for the other three models which assign literal readings for more concrete events. In comparison to the label distribution based on human annotations, there is a notable absence of implausible instances across all models. In particular, Mistral does not produce a single majority assignment for *implausible*. Similarly to the analysis of human judgements, we conduct a quantitative analysis to explore the relationship between figurative language and abstractness, original label, and PAP rating. Results are shown in Table 1, indicating associations ($p < [0.01, 0.05]$) with moderate effect size between *figurative* and *literal* labels and abstractness for all models except Qwen.

Few-shot prompts (cf. App. B, Figure 8) change Gemma results with a significant increase in implausible and unclear events. Interestingly, only marginally different results are observable for both Qwen and Mistral, which could either point to strong prediction stability across prompts or disregard of contextual information in the middle of a prompt. Lastly, Llama results change to overall more figurative events assigned. Additional quantitative inspections (cf. App. B, Table 5) underline the relation between plausible labels and abstractness with stronger associations than for zero-shot prompting.

We further explore for both zero- and few-shot settings **which prompt template leads to the strongest bias towards figurative interpretation** of the examined events. Results are shown in App. B, Table 6, indicating that prompt templates based on human instructions introduce the least bias across models.

In summary, our hypotheses are confirmed for **human-annotated** events: (i) The more concrete svo event constituents are, the more likely contextualization fails, i. e., events being judged as neither figuratively nor literally meaningful, but implausible (nonsensical). This finding underlines previous work on the influence of event abstractness

MODEL	ZERO-SHOT		FEW-SHOT	
	ACC.	τ	ACC.	τ
Gemma3-4B	0.27	0.09	0.30	0.01
Qwen3-4B	0.27	**-0.15**	0.25	**-0.16**
Mistral-7B	0.20	-0.10	0.21	-0.05
Llama3.1-8B	0.28	0.04	0.32	0.05

Table 2: Model performance for label prediction across prompt types. Predictions are aggregated across prompt templates and five model runs. We report *accuracy* (acc.) and *Kendall's* τ using human majority-assigned decisions as reference value. Bold: $p < 0.001$

on plausibility (Eichel and Schulte im Walde, 2023) and complements research on semantically anomalous vs. truly nonsensical expressions (Olsen and Padó, 2026). (ii) Confirming our hypothesis, the more abstract svo event constituents are, the more frequently plausibility is perceived, and the more probable is a figurative reading. In contrast, **LLM-predicted** results deviate from our hypothesis as we find (i+ii) a strong bias for plausibility, and specifically figurative language across categories for Qwen, Mistral and Gemma, while overall, Llama results are closer to human judgments. However, human-annotated results are only weakly mirrored with models trading implausibility for plausibility.

4.2. Qualitative Characteristics of Human- vs. LLM-Produced Contexts

We qualitatively evaluate generated contexts to assess how humans vs. models contextualize (im)plausible svo events. Across our 27 abstractness combinations, we sample up to four examples (one per label) produced by humans or models (zero-shot), yielding 97 contexts. We sample 3 example sentences per investigated event from human-generated contexts. We label one model generation per event. Events incorrectly (not) containing the original svo event are labeled *none*. We also assign *none* in case of more than one changed constituent or in case of constituent changes despite a plausible judgement. Whenever events correctly contain the original event but are semantically invalid, we assign the label *anomalous*. We follow Olsen and Padó (2026)'s labeling scheme and annotate contexts as *specific* if no self-reported indication of generic settings such as *fantasy* story is present in produced example contexts. In case of changes, we track altered event constituents.

Results for human-produced and model-generated contexts are reported in Table 3, highlighting a substantial number of specific contexts by both humans and LLMs. In comparison to humans, LLMs rarely predicted the label *neither* in which case an event constituent should be altered to enable contextualization. Nevertheless,

	H1	H2	H3	LL	QW	MI.	GE
Specific	94	92	93	51	52	27	17
Altered (s)	10	8	8	-	-	-	2
Altered (v)	6	13	13	-	-	-	-
Altered (o)	15	8	8	-	-	-	-
Anomalous	-	1	2	6	11	1	22
None	3	4	2	40	34	69	58

Table 3: Human (H) vs. LLM (LL: Llama, QW: Qwen, MI: Mistral, GE: Gemma) context patterns.

especially Mistral and Gemma frequently alter events despite a predicted figurative or literal label. Moreover, in the few cases were *neither* was assigned, models mostly fail to correctly change only one constituent *and* produce a valid sentence. Further, LLM contextualization strongly adheres to original event syntax, as highlighted by contexts to the originally and majority PAP implausible event *"license hinders ice"*. Qwen repeats the event (*"The event license hinders ice."*). Both Llama and Gemma add a single object (*"The event license hinders ice skating."*). Mistral's generation illustrates a common failure across all models: incorrect substitution despite a plausible judgement (*"The ice sculpture exhibition was hindered by the event license restrictions.")*

In comparison, human-produced contexts are based on a majority-assigned *neither* label with examples altering the subject (*"The recent sunny weather hinders ice for hockey player."*) or the object (*"The strict license hindered access to the restricted research facility."*) or providing an actually supporting, non-anomalous context (*"A license doesn't have the capability to hinder ice from forming."*) While humans use a wide vocabulary range and vary syntax to generate meaningful contexts, all models over-use the term *event*, and adhere to mostly nouns, verbs, and simple syntactic structures. In conclusion, LLM results reveal shallow patterns when compared with human-generated contexts exhibiting great nuance at assessing plausibility and contextualizing and 'repairing' events.

5. Conclusion

This work explored the connection between plausibility and (non-)literalness at the example of svo events in English. Using a carefully selected section of the PAP dataset (Eichel and Schulte im Walde, 2023), we collected and analyzed human- vs. LLM-generated judgments and examples. Our analysis reveals substantial differences between human and LLM assessments for the examined events. While humans excel at nuanced detection and contextualization of (non-)literal vs. implausible events, LLM results reveal shallow context patterns and a strong bias towards plausibility.

6. Acknowledgements

This research was supported by the Hanns Seidel Foundation's Talent Program (first author) and the DFG Research Grant SCHU 2580/4 *MUDCAT – Multimodal Dimensions and Computational Applications of Abstractness*. We also thank the reviewers for their helpful comments and nuanced feedback.

7. Limitations and Ethical Considerations

A first obvious limitation of our work is the sole focus on the English language. We expect results to differ for other languages and encourage work on plausibility, (non-)literalness, and abstractness. In this paper, we present a collection of (non-)literalness judgments and examples sentences collected via crowd-sourcing. We employ control items as well as post-processing to minimize the impact of unreliable annotations on our analyses. Approaches of mitigation could be concentrating on labels with high majorities of one label assigned or use e. g., probabilistic approaches to aggregate labels. We pay participants fairly and seek transparent communication of decisions whenever necessary during the annotation approval process. Furthermore, in our work we use a reasonable set of heuristics to parse LLM-generations. It is possible that more complex approaches might lead to different results based on the parsing process. Finally, we use LLMs that are known to exhibit bias which might be reflected in the way events are judged as well as in the style and content of generated contexts.

8. Bibliographical References

Lawrence W Barsalou and Katja Wiemer-Hastings. 2005. Situating Abstract Concepts. *Grounding cognition: The role of perception and action in memory, language, and thought*, pages 129–163.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

Harald Cramér. 1999. *Mathematical methods of statistics*, volume 9. Princeton university press.

Annerose Eichel and Sabine Schulte im Walde. 2023. A Dataset for Physical and Abstract Plausibility and Sources of Human Disagreement. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 31–45, Toronto, Canada. Association for Computational Linguistics.

Ali Emami, Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. ADEPT: An adjective-dependent plausibility task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7117–7128, Online. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, page 25–30, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models.

Bernadeta Griciūtė, Marc Tanti, and Lucia Donatelli. 2022. On the cusp of comprehensibility: Can language models distinguish between metaphors and nonsense? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 173–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

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, Miami, Florida, US. Association for Computational Linguistics.

Mohammed Khaliq, Diego Frassinelli, and Sabine Schulte Im Walde. 2024. Comparison of Image Generation Models for Abstract and Concrete Event Descriptions. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 15–21, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2026. Literally Concrete or Figuratively Abstract? Multilingual Concreteness Norms for Verb-Object Expressions. *Transactions of the Association for Computational Linguistics (TACL)*. In press.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Katrin Olsen and Sebastian Padó. 2026. Finding Sense in Nonsense with Generated Contexts: Perspectives from Humans and Language Models. Manuscript.

Karl Pearson. 1991. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Breakthroughs in Statistics: Methodology and Distribution*, page 11.

Ian Porada, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. Can a gorilla ride a camel? learning semantic plausibility from text. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 123–129, Hong Kong, China. Association for Computational Linguistics.

Ian Porada, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. Modeling event plausibility with consistent conceptual abstraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1732–1743, Online. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 Technical Report.

Philip Stuart Resnik. 1993. *Selection and information: A class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74.

A. Appendix

A.1. Data

Table 4 presents an overview of the target svo events sampled from the PAP dataset.

orig. labels	**PAP ratings**		
	plausible	*disagree*	*implausible*
plausible	81	64	31
implausible	81	80	77

Table 4: Overview of number of target event triples sampled from PAP.

A.2. Human Annotation

Human Annotator Demographics Figure 4 shows an overview of annotator demographics. Subplot (a) visualizes the age distribution of involved annotators. Mean age is 39.5 years and median age is 38 years. Subplot (b) presents the distribution between female and male participants. Please note that this is based on self-reported biological sex of participants. We do not collect information on gender identity. We report annotators' employment status in subplot (c) with the majority of participants either working full-time or part-time. "No paid work" refers to individuals focusing on care work as well as retired or disabled individuals. "Soon new job" denotes participants who start a new job in the next month (which does not mean that they are not employed at the moment they took part in the study). "Unemployed" implies that someone is unemployed *and* job-seeking. "DATA_EXPIRED" refers to long-time participants on Prolific who have not updated their Prolific profile for a longer period of time. Some information such as employment or student status hence might get marked as expired. As visualized in subplot (d), self-reported simplified ethnicity groups are mainly White (67%) and Black (25%). While nationality as shown in subplot (e) needs to include UK, Ireland, U.S., or Australia, participants might have dual citizenship (e.g., the UK allows for that). Subplot (f) lists countries where participants reside.

A.3. Modeling

We present prompt templates for zero- and few-shot prompting in Figure 6 and Figure 6.

B. Results

Human vs. Model Label Predictions We present **zero-shot** model results for predicting figurative labels for all models in Figure 7. **Few-shot**

Guidelines

You will be given 30 three-word events such as "cat eat sardine" or "friend grasp meaning".

Tasks:

1. Decide whether the event description is clearly based on the meanings of the three words.
For example**, the event "cat eat sardine" is literally describing the event of a cat eating a sardine**. In contrast, **the event "friend grasp meaning" does not describe a friend literally grasping a meaning: the event meaning is figurative,** rather than literal.

2. Together with your decision regarding whether an event is literal or figurative, we also ask you to **provide an example sentence** including the three-word event. For example:

- Literal usage: *I saw a cat eating a sardine near the lake today.*
- Figurative usage: *My friend quickly grasped the meaning of the mathematical problem.*

3. In some cases you will neither be able to identify a literal nor a figurative meaning, because the event is absolutely implausible, and it is not possible to come up with any interpretation at all. An example of such an **absolutely implausible event is "plea overrun rain"**. In this case, we ask you to alter one of the three words and then **provide an example sentence**. For instance, changing "*overrun*" to "*summons*" forms the phrase "**plea summons rain**".

- Adjusted event: *The farmer's plea summons the rain to save the crops.*

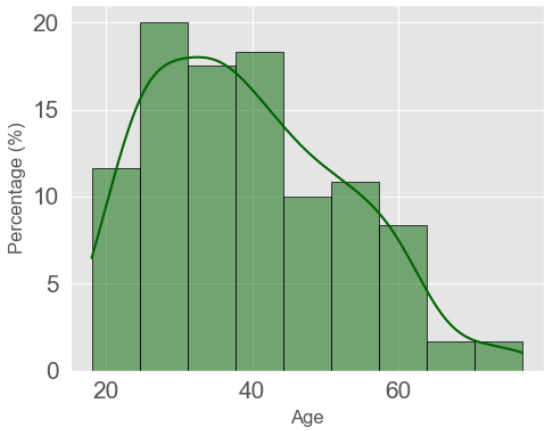
Decision for "reign spreads power" *

- ○ literal
- ○ figurative
- ○ neither: the event is implausible

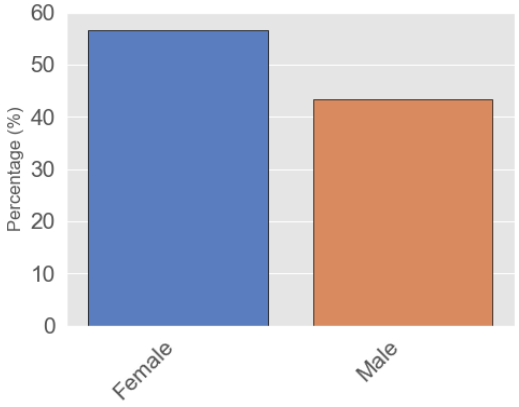
Example sentence for "reign spreads power" (or an event variation, if this event is absolutely implausible)

Your answer

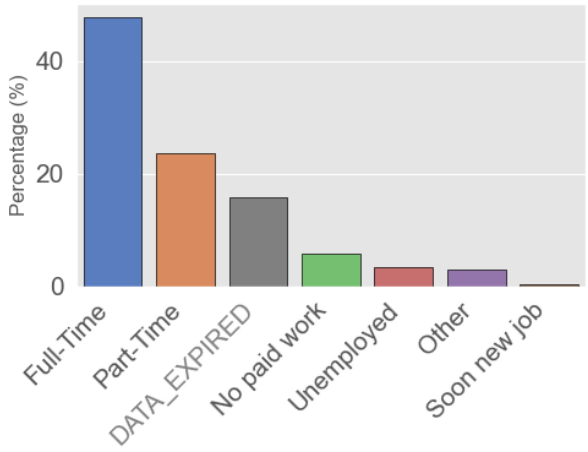
Figure 3: Annotation interface



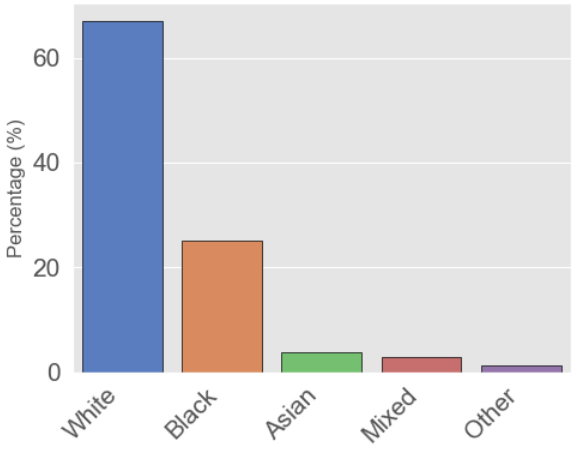
(a) Age.



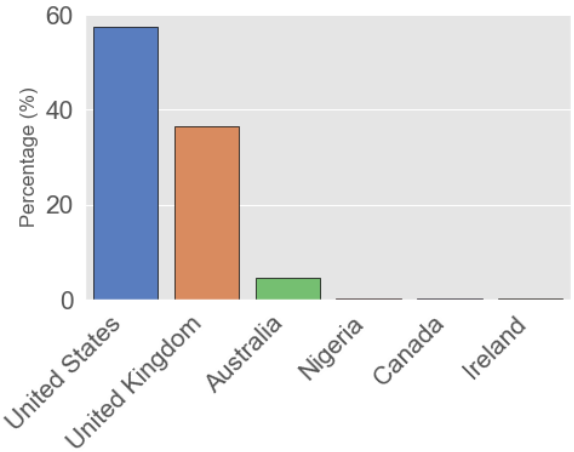
(b) Female and male (sex; gender identity information not collected).



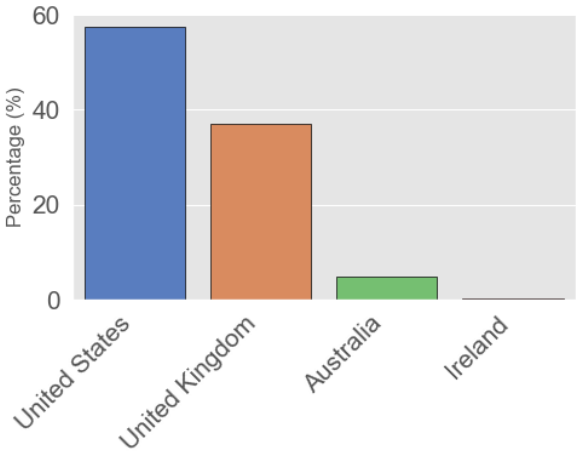
(c) Employment status.



(d) Self-reported simplified ethnicity groups.



(e) Nationalities (dual citizenship possible).



(f) Country of residence.

Figure 4: Distributions of annotator demographic features.

model results for predicting figurative labels are presented for all models in Figure 8.

Further, quantitative analysis results for few-shot modeling results are presented in Table 5. We include values from human analysis for reference.

We analyze which prompt template leads to strongest bias towards figurative interpretation. We consider both zero- and few-shot prompt templates and show results in Table 6.

```
Is the following event figurative or literal or neither?
Event: {event}
Answer with only one label: figurative or literal or neither.
Respond in the following format:
Label: <figurative|literal|neither>
```

```
Determine whether the event below is figurative or literal or neither.
Event: {event}
Respond in the following format:
Label: <figurative|literal|neither>
```

```
Decide whether the event description is clearly based on the meanings of the three
↪  words.
Event: {event}
Answer with only one label: <figurative|literal|neither>.
Decision: <label>
```

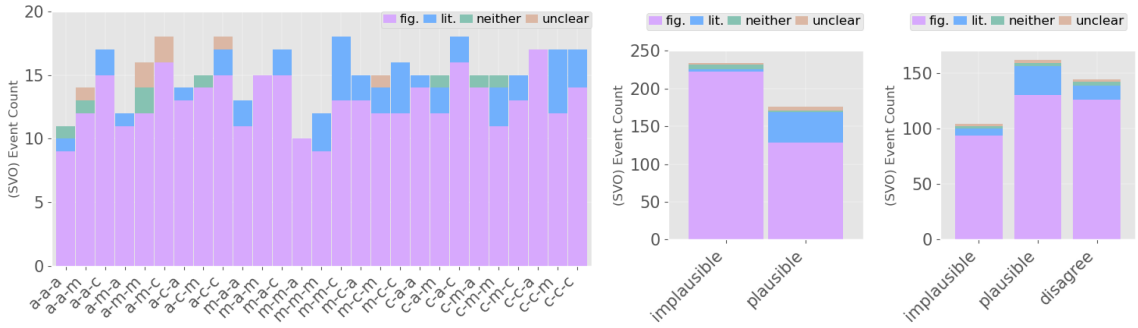
Figure 5: **Zero shot** prompt templates, from top to bottom including the task formulation as a *question* (top), a statement (middle), and as closely as possible based on the instruction for human annotation but condensed in one sentence (bottom).

```
Decide whether the event description is clearly based on the meanings of the three
↪  words.
For example, the event "cat eat sardine" is literally describing the event of a cat
↪  eating a sardine.
In contrast, the event "friend grasp meaning" does not describe a friend literally
↪  grasping a meaning:
the event meaning is figurative, rather than literal.
Event: {event}
Answer with only one label: <figurative|literal|neither>
Decision: <label>
```

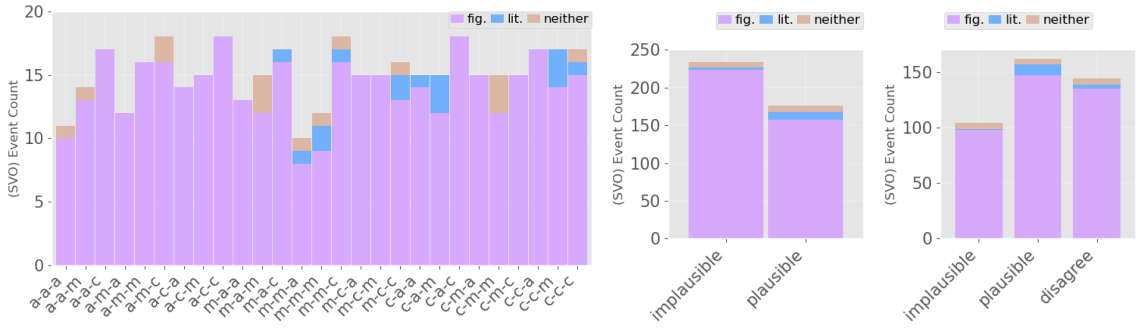
Figure 6: **Few shot** prompt template. The prompt is based as closely as possible on the instruction for human annotation while optimizing for the shortest length.

		Human		**Qwen3-4B**		**Llama3.1-8B**		**Mistral-7B**		**Gemma3-4B**	
Figurative	df	χ^2	V	χ^2	V	χ^2	V	χ^2	V	χ^2	V
ABSTRACTNESS	26	**63.64*** **	0.39	43.66*	0.33	46.71**	0.34	51.37**	0.35	53.59**	0.36
ORIG. LABEL	1	6.91**	0.13	41.05	0.32	7.74**	0.14	4.64*	0.11	**13.56*** **	0.18
PAP_RATING	3	13.49**	0.18	15.48**	0.19	4.04	0.10	3.37	0.09	1.23	0.05
Literal											
ABSTRACTNESS	26	43.93**	0.33	50.84**	0.35	**61.28*** **	0.39	34.08	0.29	**57.32*** **	0.37
ORIG. LABEL	1	**111.33*** **	0.52	40.70	0.32	**25.35*** **	0.25	10.02**	0.16	**19.26*** **	0.22
PAP_RATING	3	**17.29*** **	0.21	15.73**	0.2	7.37	0.13	3.91	0.10	1.97	0.07
Neither											
ABSTRACTNESS	26	32.80	0.28	33.25	0.28	23.17	0.24	33.14	0.28	29.15	0.27
ORIG. LABEL	1	71.96	0.42	0.02	0.01	0	0	-	-	6.45*	0.13
PAP_RATING	3	13.73**	0.18	1.54	0.06	1.86	0.07	6.54	0.13	8.39*	0.14
Unclear											
ABSTRACTNESS	26	29.59	0.27	28.55	0.26	20.12	0.22	51.42**	0.35	39.34*	0.31
ORIG. LABEL	1	8.23**	0.14	0	0	1.02	0.05	0	0	0.87	0.05
PAP_RATING	3	4.17	0.10	3.73	0.10	10.21*	0.16	1.89	0.07	9.44**	0.15

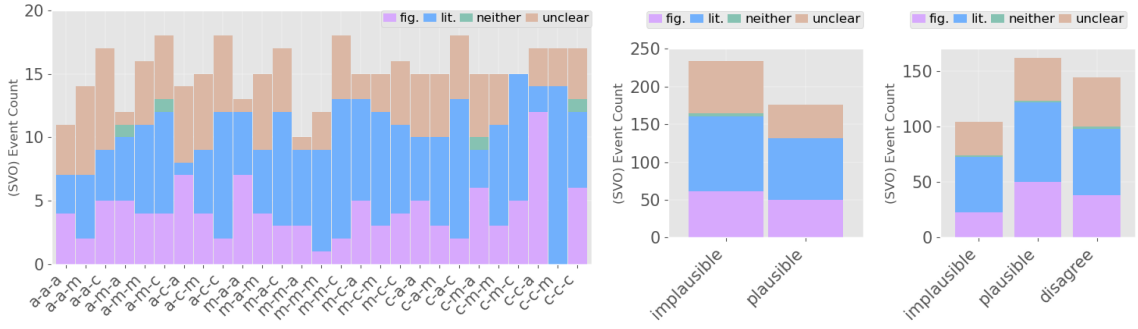
Table 5: Associations between figurative language and abstractness, original label, or PAP ratings. χ^2 indicates *significance* ($p < 0.05$:**, $p < 0.01$:**, $p < 0.001$: ***) and Cramér's V measures *strength* of association. Model results are based on **few-shot** prompts.



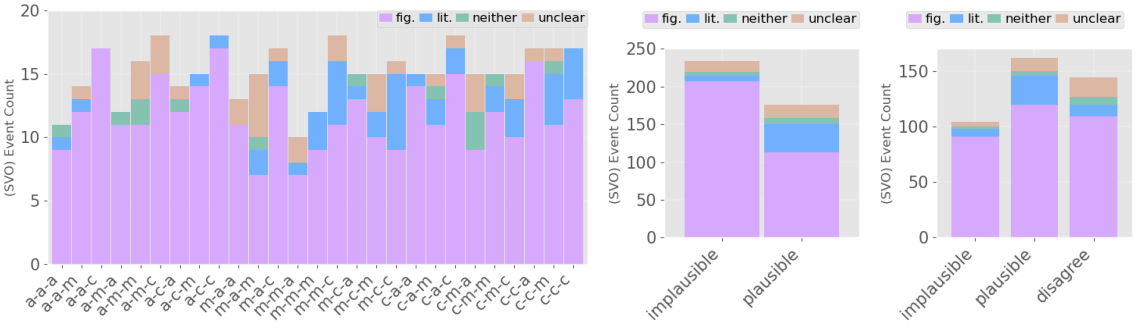
(1a) **`Qwen3-4B`** judgements across abstractness comb. (1b) orig. label (1c) PAP rating



(2a) **`Mistral-7B`** judgements across abstractness comb. (2b) orig. label (2c) PAP rating

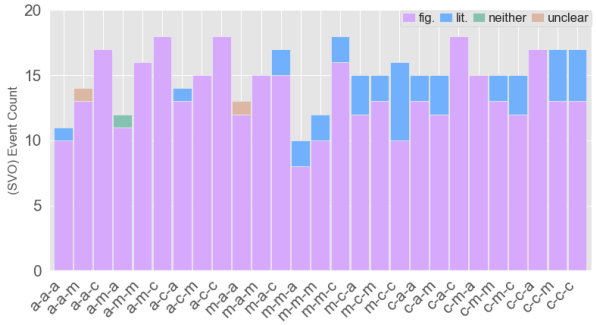


(3a) **`Llama3.1-8B`** judgements across abstract. comb. (3b) orig. label (3c) PAP rating

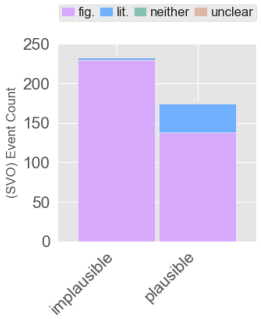


(4a) **`Gemma3-4B`** judgements across abstract. comb. (4b) orig. label (4c) PAP rating

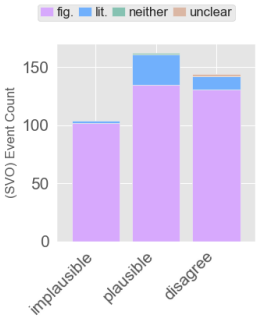
Figure 7: Overview of analysis of **`Qwen3-4B`**, **`Mistral-7B`**, **`Llama3.1-8B`**, and **`Gemma3-4B`** figurative labels (**zero-shot**), similarly to human analysis in Figure 1.



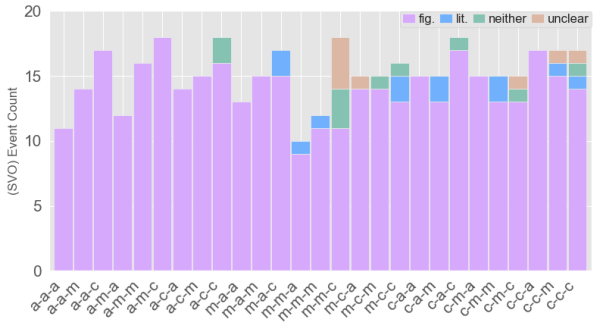
(1a) **`Qwen3-4B`** judgements across abstractness comb.



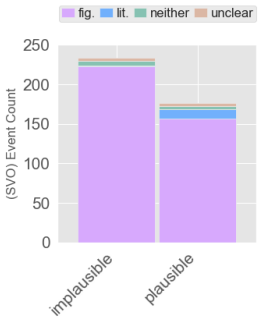
(1b) orig. label



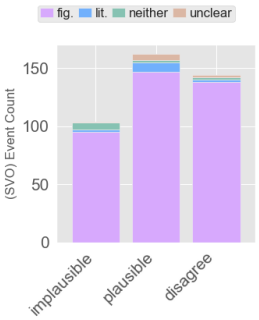
(1c) PAP rating



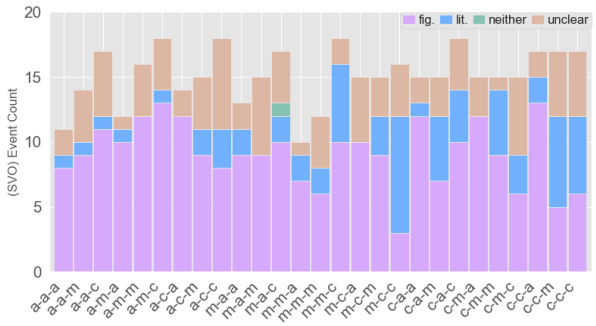
(2a) **`Mistral-7B`** judgements across abstractness comb.



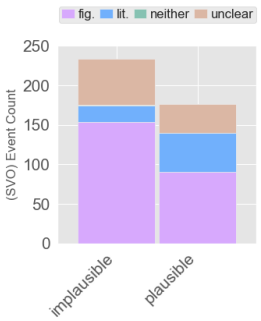
(2b) orig. label



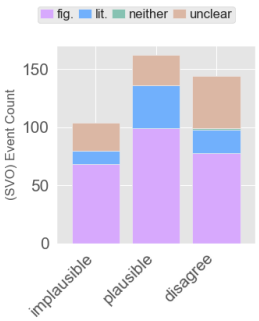
(2c) PAP rating



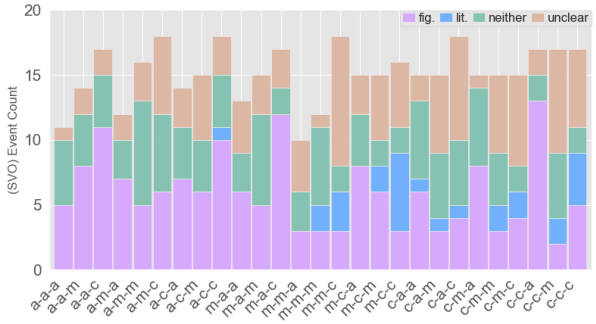
(3a) **`Llama3.1-8B`** judgements across abstract. comb.



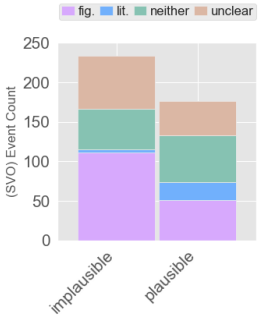
(3b) orig. label



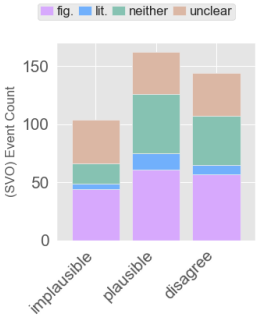
(3c) PAP rating



(4a) **`Gemma3-4B`** judgements across abstract. comb.



(4b) orig. label



(4c) PAP rating

Figure 8: Overview of analysis of **`Qwen3-4B`**, **`Mistral-7B`**, **`Llama3.1-8B`**, and **`Gemma3-4B`** figurative labels (**few-shot**), similarly to human analysis in Figure 1.

	Zero-shot			Few-shot
	question	statement	human instr.	human instr.
Gemma3-4B	67.88	89.29	61.05	22.83
Qwen3-4B	91.00	83.94	63.50	73.90
Mistral-7B	84.91	95.83	78.47	94.51
Llama3.1-8B	23.88	31.39	31.03	61.37

Table 6: Overview of share of only figurative labels per prompt, in percent. Note that percentages do not add up to 100% but each number is a share of 100% labels distributed among *figurative* (shown here), *literal*, *neither*, and *unclear* (not shown). For reference, prompt templates are listed in Figure 6 and Figure 5.