# PAP: A Dataset for Physical and Abstract Plausibility and Sources of Human Disagreement

Annerose Eichel and Sabine Schulte im Walde

Institute for Natural Language Processing, University of Stuttgart, Germany

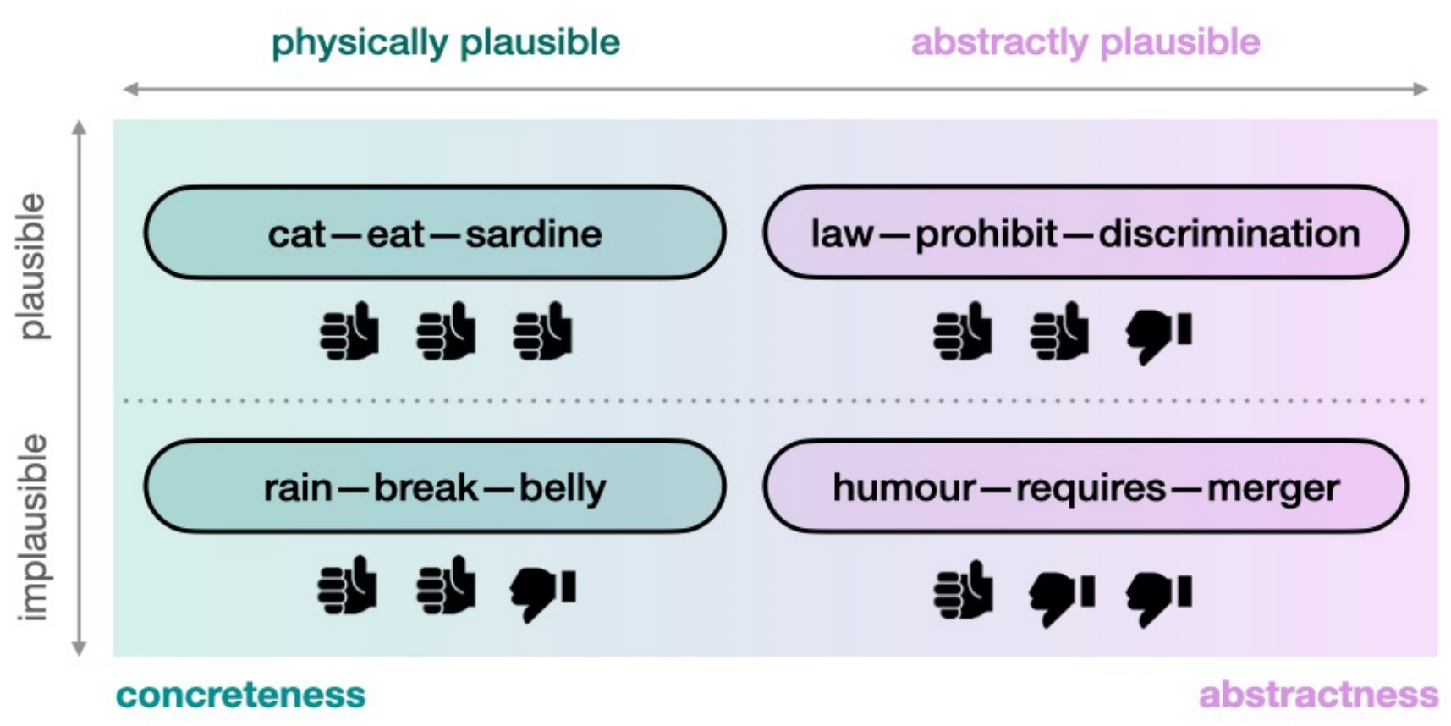
{annerose.eichel,schulte}@ims.uni-stuttgart.de

Universität Stuttgart

LAW-XVII
2023

## INTRODUCTION

- Discerning plausible from implausible events: crucial building block for NLP
- Previous work mostly focused on semantic knowledge to distinguish
  - *__physically__* plausible vs. implausible events
  - events with mostly conceptually ***concrete*** participants



### Research Goals & Contributions

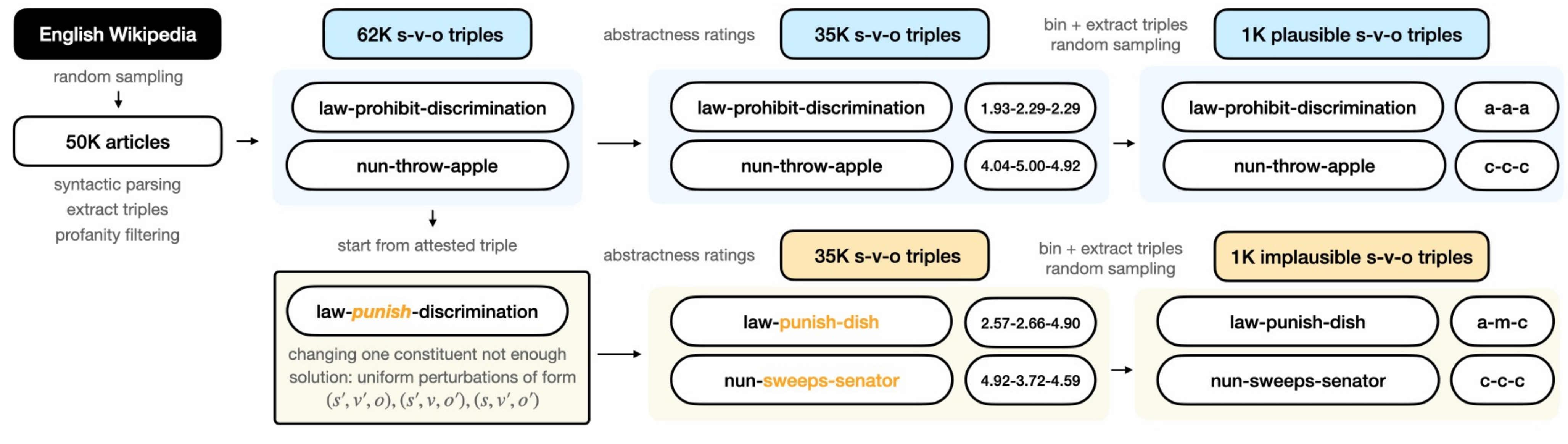
- Create novel dataset for **physical and abstract plausibility** of events in English, capturing abstractness to the same extent as concreteness for the first time
- Systematically examine **plausibility across levels of abstractness**
- Explore and represent **disagreement in plausibility annotation**

## CAPTURING (SEMANTIC) PLAUSIBILITY

### Plausibility

- Captures non-surprisal in a given context
  ***child-sleep*** vs. ***tree-sleep***
- Includes both what is preferred (and probably most plausible) and what is unusual (but still very much plausible),
  ***child-eat-banana*** vs. ***child-eat-pebble***
  → in contrast to selectional preference / thematic fit
- Can be estimated as a matter of degree with events assessed corresponding to perceived plausibility
  ***child-eat-banana*** vs. ***child-eat-pebble*** vs. ***child-eat-skyscraper***
- Denotes what is likely in a given world but not necessarily attested in a given corpus
  ***human-dies*** vs. ***human-breathes***

## CONSTRUCTING EVENT TARGETS



Simplified Illustration of Dataset Construction

### Plausible Events (marked in blue)

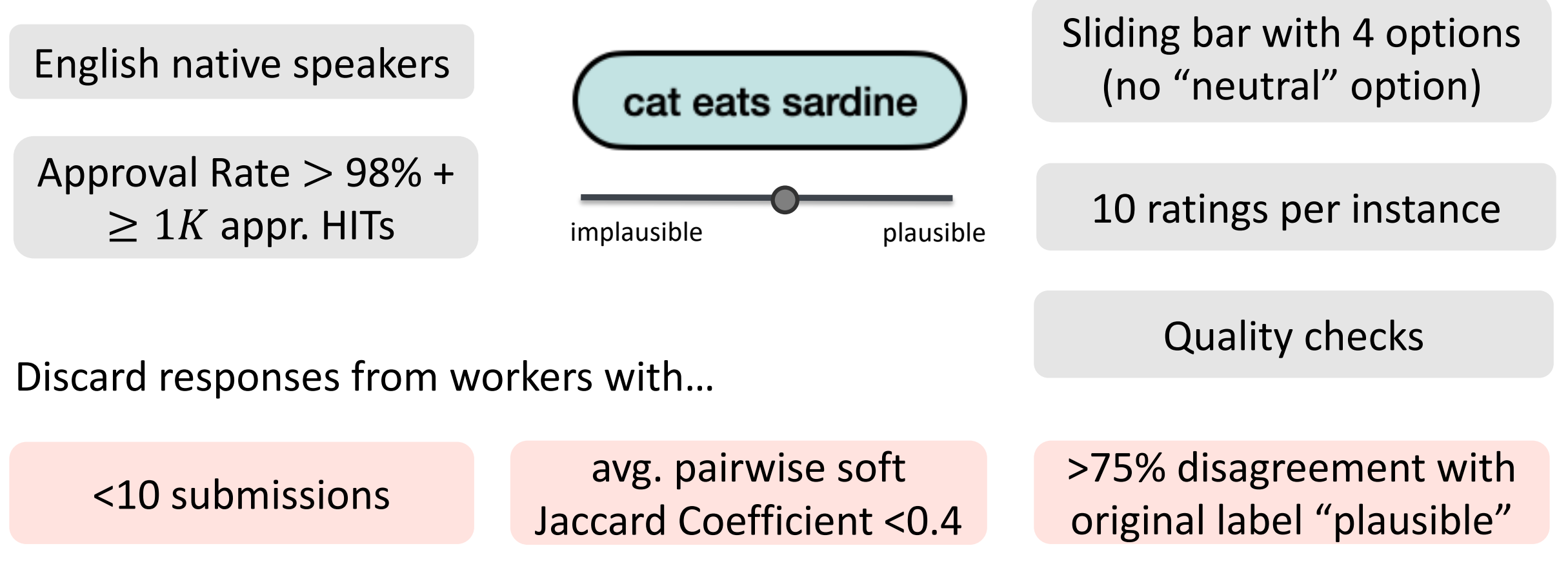
- From English Wikipedia sample:
  Extract attested triples, filter for profanity, assign abstractness ratings, bin according to abstractness, and sample 1,080 plausible event triples for 27 abstractness combinations

### (Pseudo-)Implausible Events (marked in yellow)

- Based on extracted attested triples:
  (i) Automatically generate pseudo-implausible triples by perturbating triple constituents
  (ii) Construct 1,080 pseudo-implausible triples similarly to plausible triple construction

## COLLECTING HUMAN ANNOTATIONS

**Task:** Collect plausibility judgements on AMT for 2,160 plausible and implausible triples



English native speakers

Approval Rate $>$ 98% + $\geq 1K$ appr. HITs

Sliding bar with 4 options (no "neutral" option)

10 ratings per instance

Quality checks

Discard responses from workers with...

<10 submissions

avg. pairwise soft Jaccard Coefficient <0.4

>75% disagreement with original label "plausible"

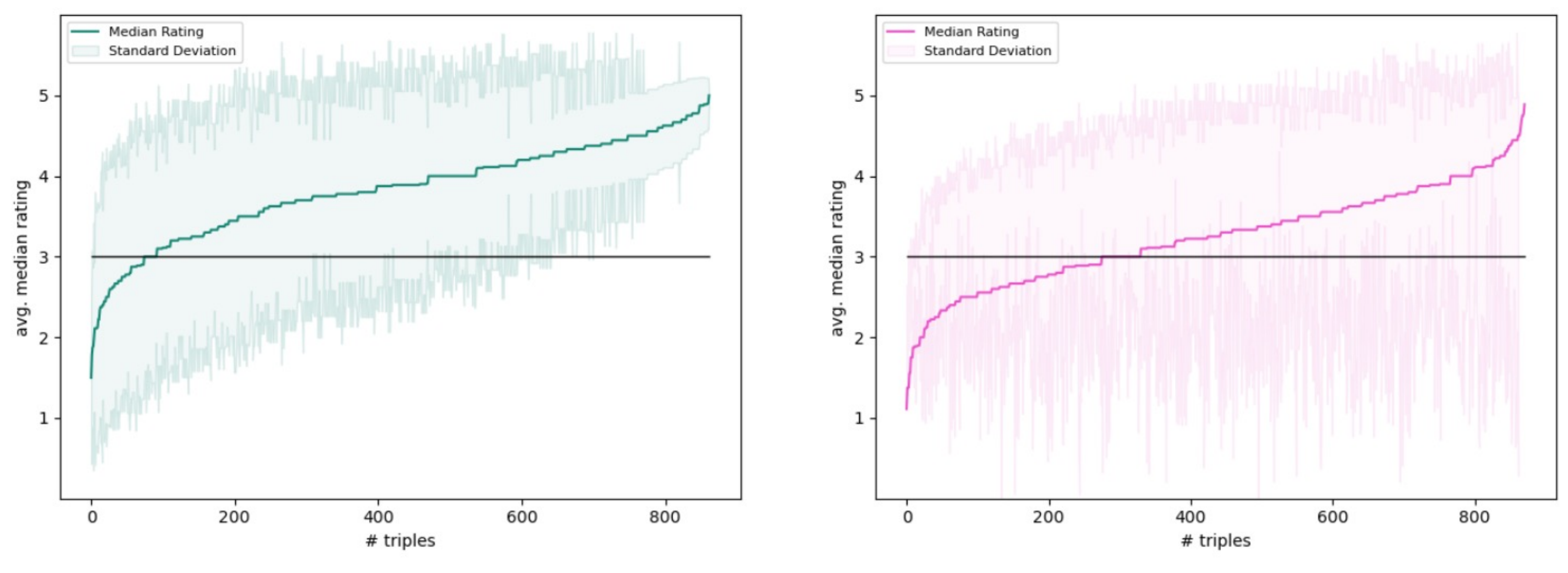
Only keep triples with min. 8 ratings

### Dataset Statistics

- **15,571 plausibility ratings for 1,733 triples**
- Ø 8.9 ratings per triple
- Ø 32 triples per abstractness combination
- IAA: Soft Jaccard Coefficient of 0.64
  → reasonable agreement among annotators with indication of disagreement to be examined

## ANALYSIS OF HUMAN JUDGEMENTS AND DISAGREEMENT

### What can we learn from rating distributions?



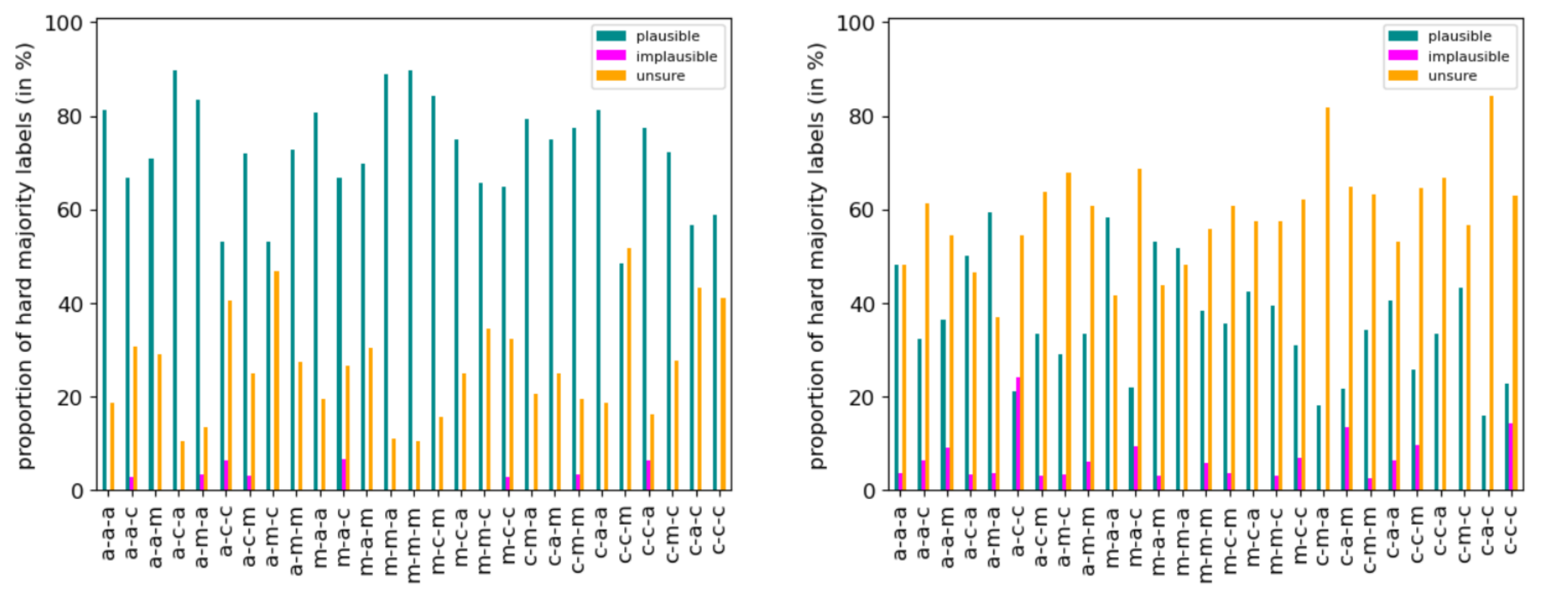
(a) Average median rating across plausible triples.

(b) Average median rating across implausible triples.

Figure 4: Average median ratings across originally plausible (a) and implausible (b) triples with standard deviation visualized as cloud around average rating lines. Triples are represented numerically on the x-axis. The black horizontal line denotes a median rating of 3. Average median ratings for *plausible* triples *below* the line disagree with the original label, while the opposite is true for average median ratings for *implausible* triples. Here, ratings *above* the line disagree with the original label.

(i) Humans tend to favor plausibility over implausibility, while avoiding the extreme on the plausibility end of the scale.
(ii) Implausibility yields higher disagreement as annotators disagree more when rating triples originally labelled implausible.

### How does abstractness impact plausibility ratings?



(a) Attested plausible triples.

(b) Perturbed implausible triples.

Figure 5: Proportion of strict majority ratings (⩾70%) across abstractness combinations for attested plausible triples (a) and perturbed implausible triples (b). Green bars denote a majority of plausible ratings $\in \{4, 5\}$, pink bars refer to a majority of implausible ratings $\in \{1, 2\}$, and orange bars capture cases of no clear majority.

(i) Plausibility tends to be more likely to be assigned in case of more abstract event participants.
(ii) Implausibility seems to be easier to capture with conceptually concrete words – no matter the underlying original label.

## ACKNOWLEDGEMENTS

Hanns Seidel Stiftung

Institut für Maschinelle Sprachverarbeitung

## CONCLUSIONS

- Presented a novel human-annotated dataset for physical and abstract plausibility for events in English
- Explored relationship between abstractness and plausibility and analyzed annotator disagreement
- Released both raw and a range of aggregated annotations to foster research on (semantic) plausibility and related notions, disagreement, and relevant downstream tasks such as commonsense reasoning

Scan and use PAP