When Choosing Plausible Alternatives, Clever Hans can be Clever

Pride Kavumba Naoya Inoue Benjamin Heinzerling
Keshav Singh Paul Reisert Kentaro Inui

https://balanced-copa.github.io

Clever Hans performed arithmetic by exploiting cues from handlers

Clever Hans Effect in NLP

NLI: models perform well with incomplete input [Gururangan+18; Poliak+18; Dasgupta+18]

Machine Reading Comprehension: superficial cues make questions easier [Sugawara+18]

Argument Reasoning Comprehension: BERT exploits superficial cues (e.g. *not*). Nearly random performance without cues [Niven+19]

COPA: Choice Of Plausible Alternatives [Roemmele+11]

Benchmark for causal reasoning

Part of SuperGLUE [Wang+19]

Example:

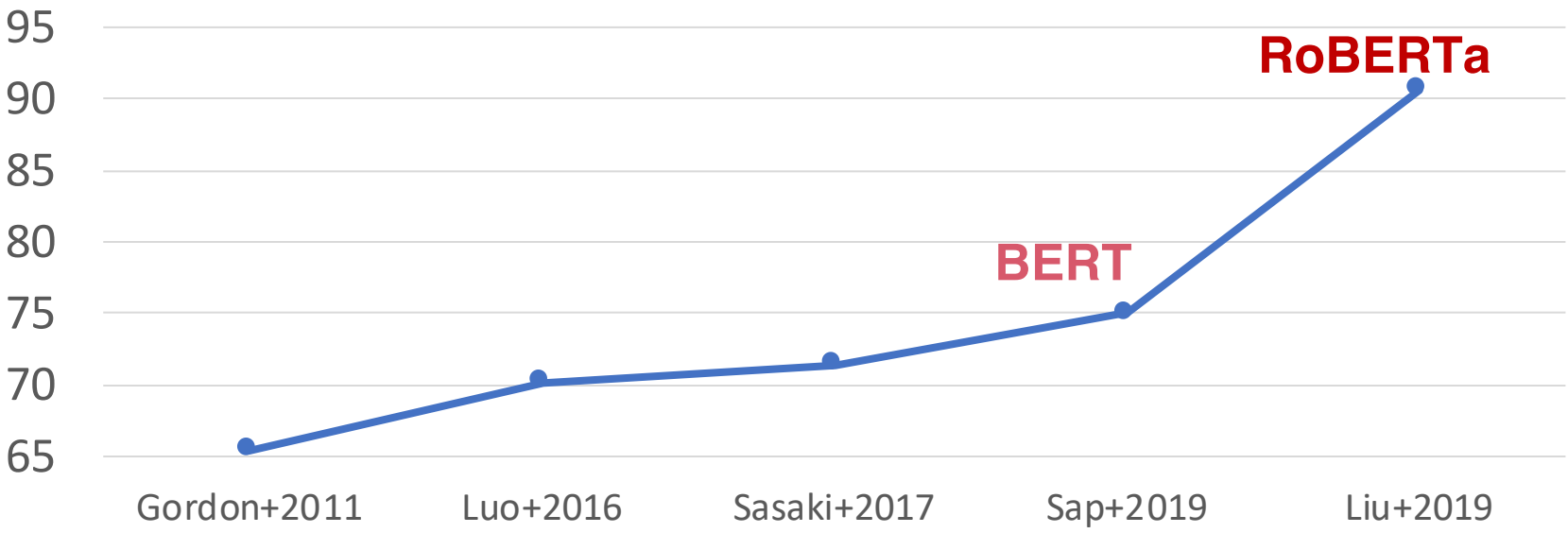
Premise: The woman hummed to herself.

Question: What was the cause for this?

Alternative1: She was in a good mood. ✓

Alternative2: She was nervous.

Rise of the Muppets on COPA



Is this the Clever Hans effect?

Research Questions

1. Does COPA have superficial cues?
2. If so, do pre-trained language models exploit these cues?
3. If they do, how do LMs perform without cues?

Superficial Cues in COPA

Superficial cues:

- Uneven token distributions across classes
- Allow models to use simple heuristics to solve

We found cues in COPA:

- Some tokens appear more often in one alternative
- Most informative cues: **in, was, to, the, a**

These cues are predictive of the correct choice

Research Questions

1. Does COPA have superficial cues?
   Yes! ☹

2. Do pre-trained language models exploit these cues?

RoBERTa [Liu+19] exploits superficial cues

Experiment: provide only incomplete input
Makes the task impossible

Question: What was the cause of this?

A1: She was in a good mood.

A2: She was nervous

RoBERTa performs better (**59.6%**) than random chance

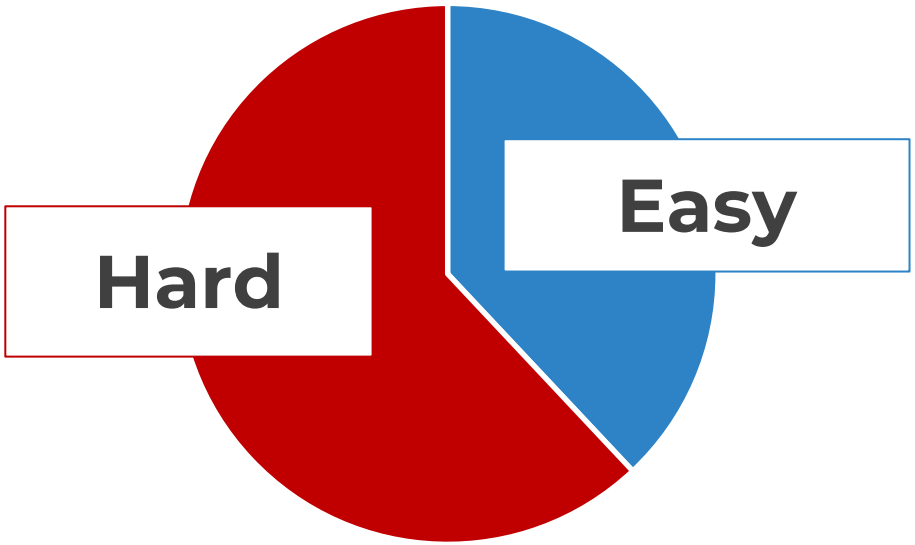
Problematic: COPA is designed as a choice between alternatives given the premise.

Splitting COPA into Easy and Hard Subsets

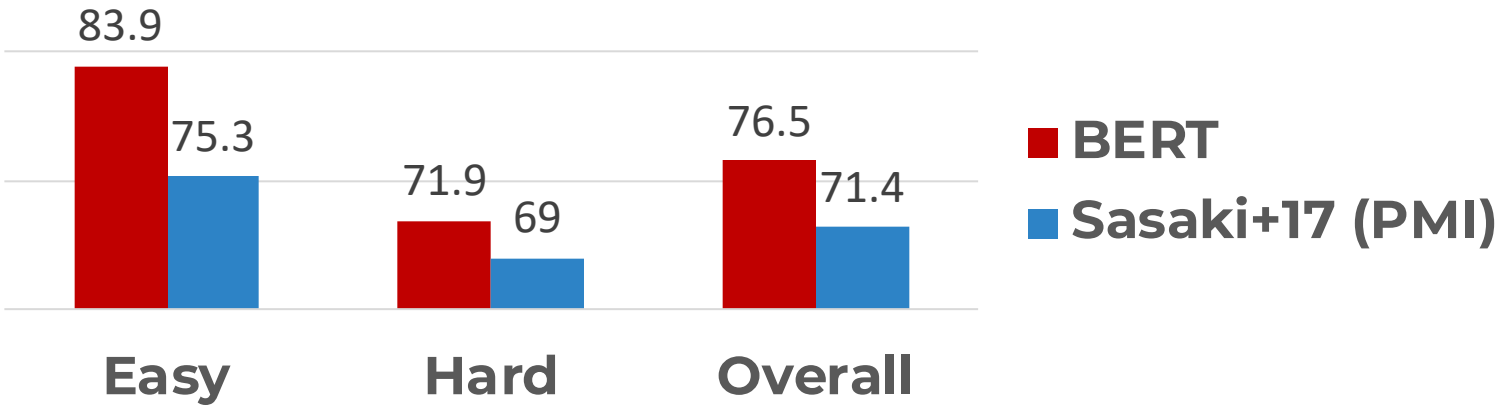
Easy subset:
instances correctly solved by RoBERTa when shown alternatives only.

Hard subset:

Remaining instances

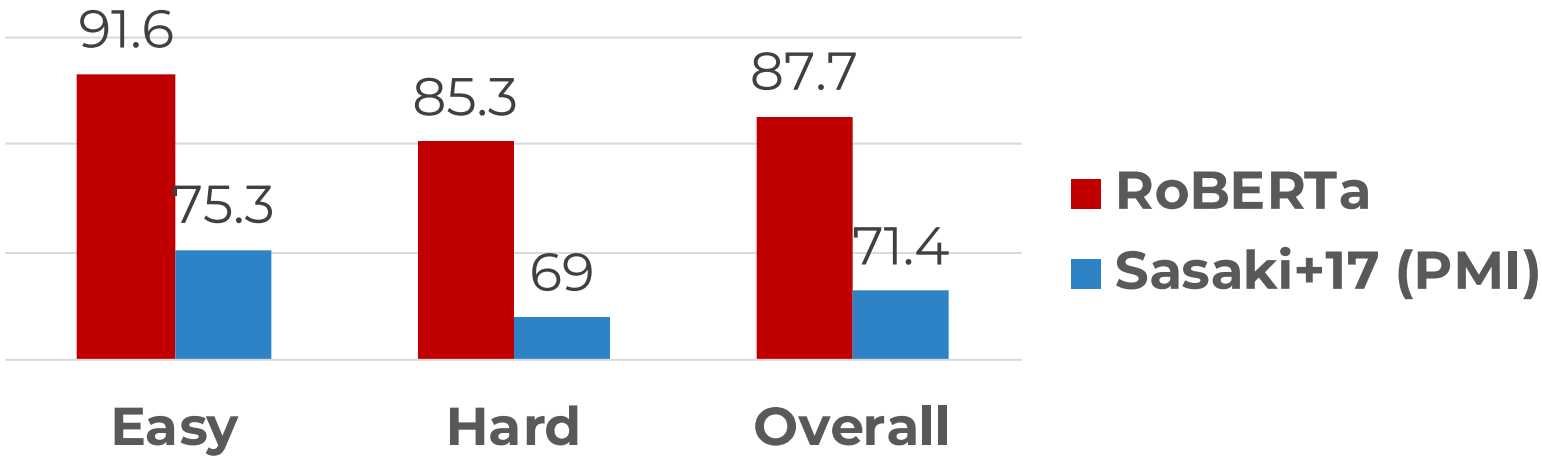


BERT on COPA



- BERT strongly exploits superficial cues
- BERT improves mainly on the Easy subset

RoBERTa on COPA



- RoBERTa also exploits superficial cues
- But: RoBERTa seems to rely less on superficial cues than BERT

Research Questions

1. Does COPA have superficial cues?

   Yes! ☹

2. Do pre-trained language models exploit these cues?

   Yes! ☹

3. How do LMs perform without cues?

Let’s fix COPA!

Balanced COPA

Balanced token distribution across alternatives
This neutralizes superficial cues

Original COPA instance

P: The woman hummed to herself.

Q: CAUSE?

✓ She was in a good mood.

She was nervous.

Mirrored COPA instance

P: The woman trembled.

Q: CAUSE?

She was in a good mood.

✓ She was nervous.

Balanced COPA

Available at

https://balanced-copa.github.io

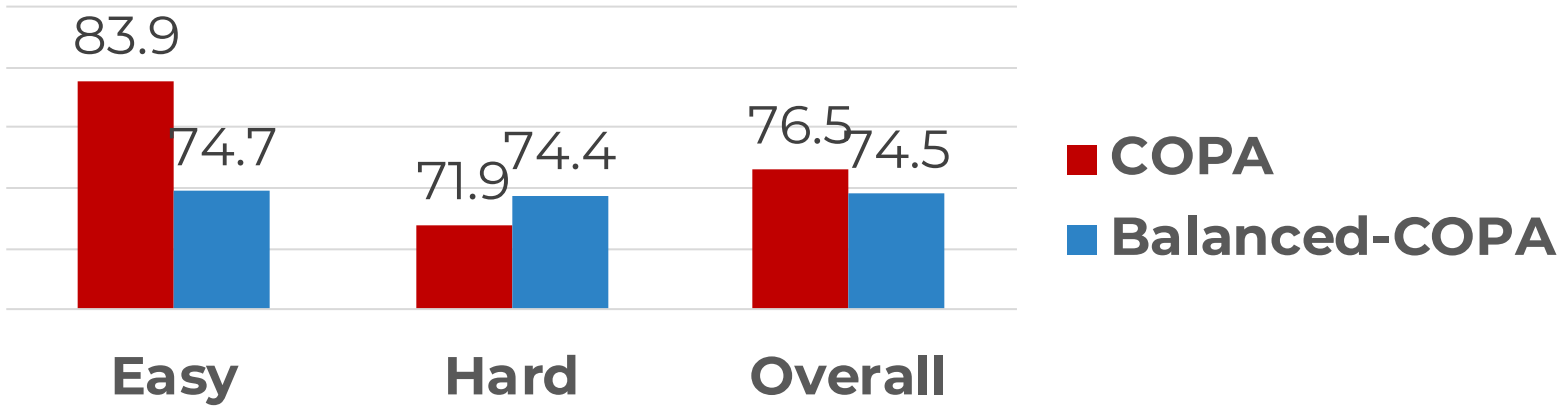
Makes superficial cues ineffective

Human Evaluation shows it is of similar quality as the original COPA

Training without Superficial Cues

- Train BERT and RoBERTa on Balanced COPA
- Test on original COPA

Balanced COPA: BERT



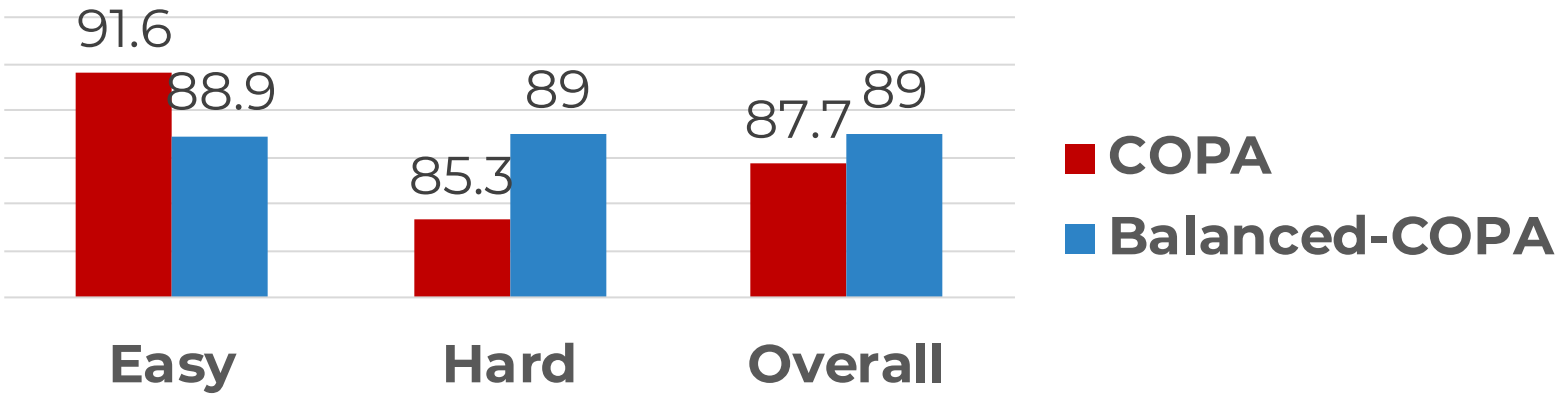
Performance on Easy subset drops significantly

- BERT strongly relied on superficial cues

Surprising improvements on Hard subset

- BERT learns the task to some degree

Balanced COPA: RoBERTa



Smaller drop on Easy subset

- RoBERTa relies less on cues

Larger improvement on Hard subset

- RoBERTa learns the task to a greater degree

Conclusions

COPA contains superficial cues

BERT exploits these cues

RoBERTa relies less on cues

Balanced COPA does not contain superficial cues (hopefully)
https://balanced-copa.github.io

Trained on Balanced COPA, BERT and RoBERTa perform well

Crowdworkers