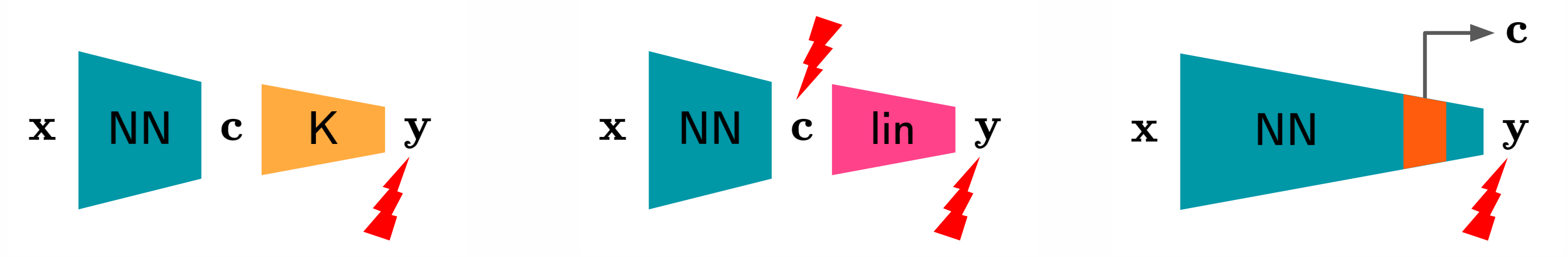
A Neuro-Symbolic Benchmark Suite for Concept Quality and Reasoning Shortcuts

S. Bortolotti[1] **E. Marconato**[1,2] T. Carraro[4, 5] P. Morettin[1] E. v. Krieken[3] A. Vergari[3] S. Teso [1] A. Passerini[1]

[1]University of Trento [2]University of Pisa [3]University of Edinburgh [4]Fondazione Bruno Kessler [5]University of Padova

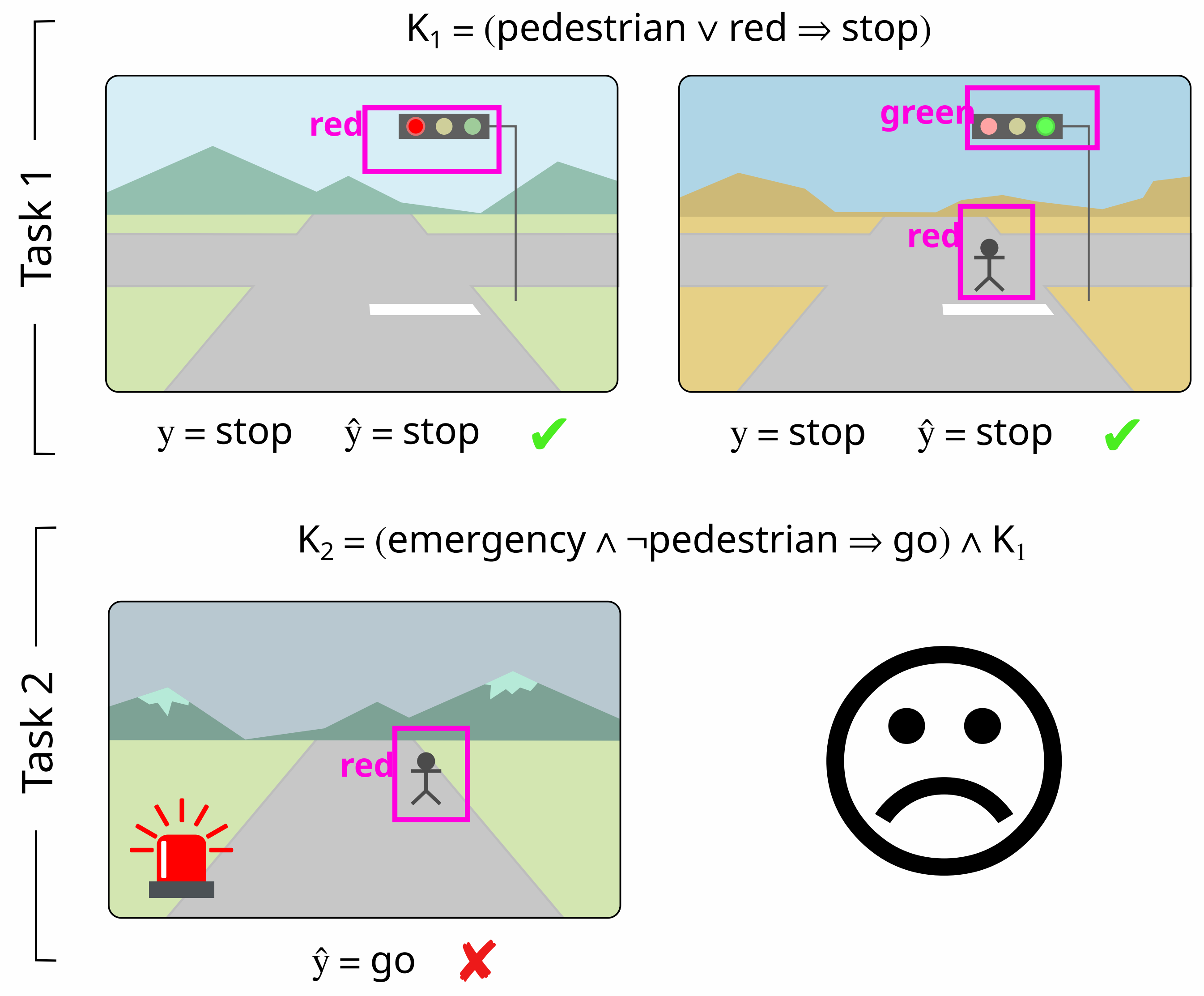
REASONING SHORTCUTS

Goal: Study supervised models that classify samples *correctly* but for the *wrong concepts*.



(a) NeSy (DPL & LTN) (b) CBM (c) NN & CLIP

Reasoning Shorcuts [1]: NeSy predictors [2], Concept-based Models [3] and VLMs like CLIP [4] solve Learning & Reasoning tasks by exploiting semantically misleading concepts.


K_1 = (pedestrian ∨ red ⇒ stop)
Task 1
red
green
red
y = stop ŷ = stop ✓
y = stop ŷ = stop ✓
K_2 = (emergency ∧ ¬pedestrian ⇒ go) ∧ K_1
Task 2
red
ŷ = go ✗

L&R TASKS

	Task	Data			Properties		
		Gen	OOD	Cont	Cplx x	Cplx K	Amb K
+/×	MNMath (new)	✓	✓	✓	✗	✓	✗
	MNAdd-Half	✗	✓✓	✗	✗	✗	–
	MNAdd-EvenOdd	✗	✓✓	✓✓	✗	✗	–
∧/∨	MNLogic (new)	✓	✓	✓	✗	✓	✗
	Kand-Logic	✓	✓	✓	✗	✓	✓
	CLE4EVR	✓	✓	✓✓	✓	✗	✓
⚠	BDD-OIA	✗	✗	✗	✓	✓	✓
	SDD-OIA (new)	✓	✓✓	✓	✓	✓	✓

FEATURES

① **Challenging**: the # of RSs can be chosen **a priori** and counted using `countrss`, allows to control task difficulty.

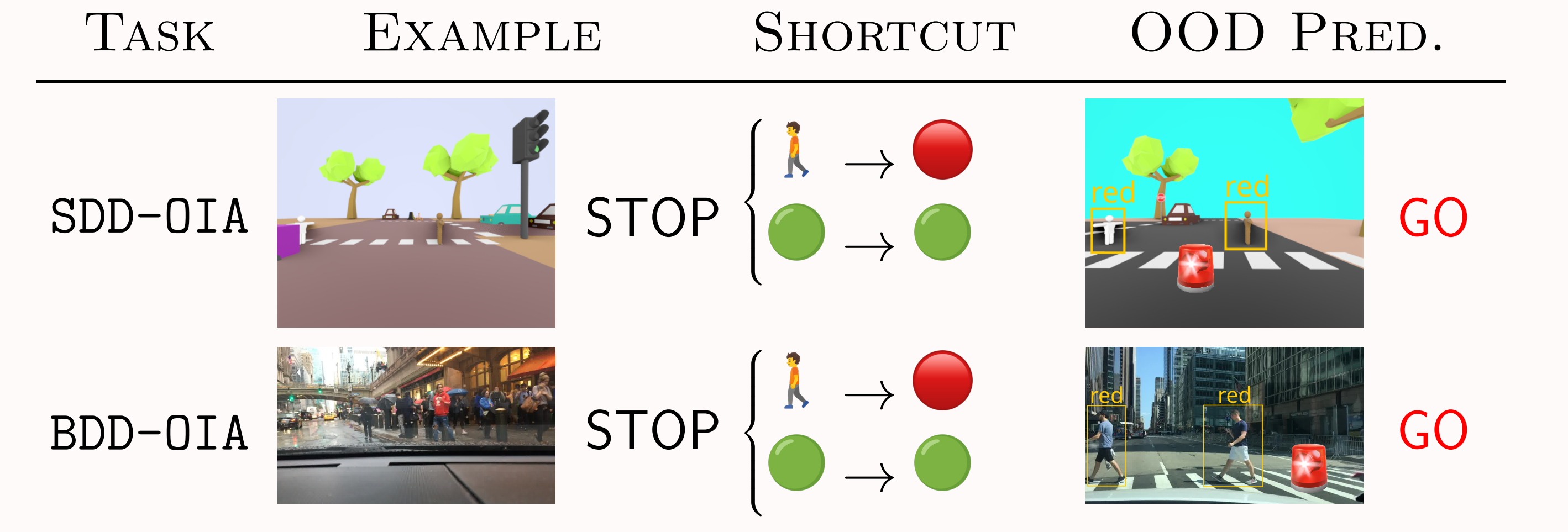
② **Configurable**: data sets & generators can be easily configured with YAML/JSON files.

③ **Intuitive**: straightforward to use:

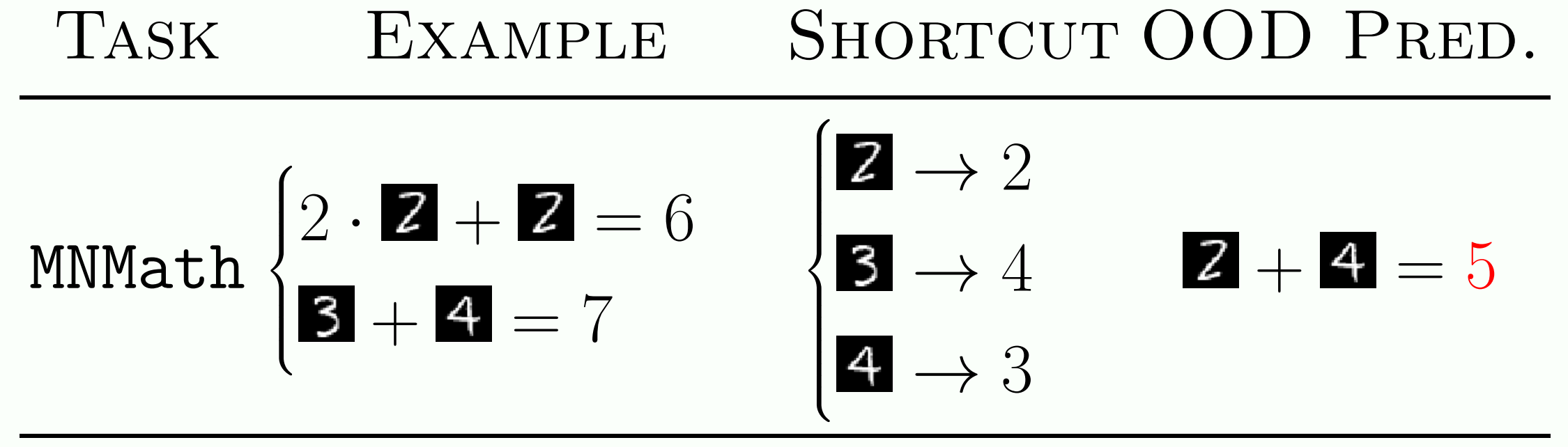
```
from rsbench import MNLOGIC

dataset = MNLOGIC(args)
train(model, dataset)
test(model, dataset)
```

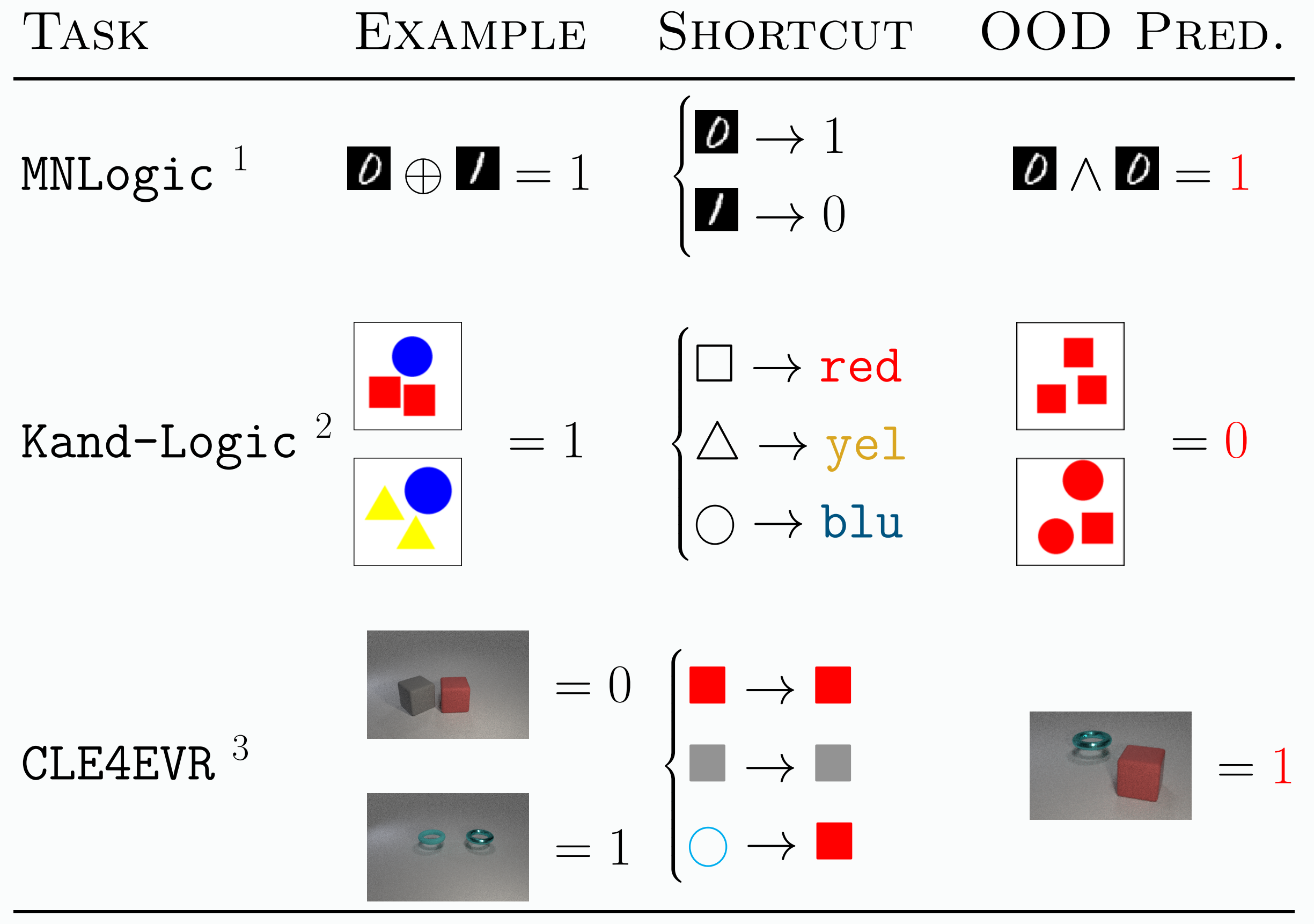
EXAMPLES



Knowledge K = *the traffic laws.*



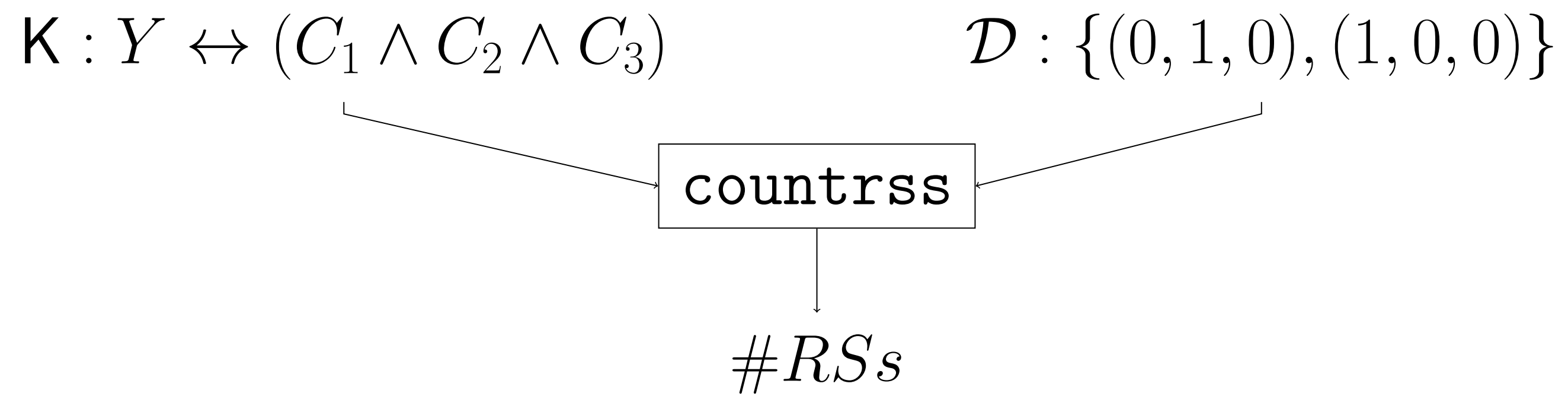
Knowledge K = *equations must hold.*



1: *Knowledge* K = *formula must hold.*
2: *Knowledge* K = *pattern must hold.*
3: *Knowledge* K = *same color and shape?*

ASSESSING RS

① **Task-level**: `countrss` counts the # of potential RSs in any L&R task!



K : $Y \leftrightarrow (C_1 \wedge C_2 \wedge C_3)$

$\mathcal{D}: \{(0,1,0),(1,0,0)\}$

countrss

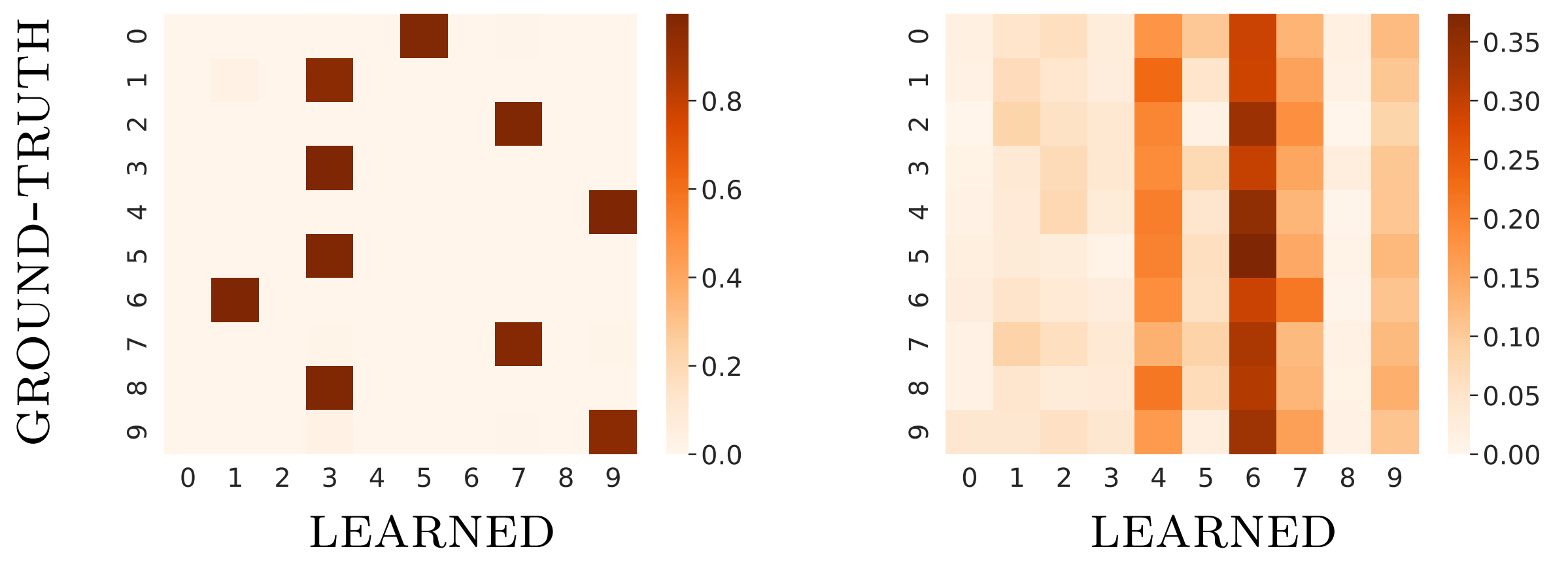
$\#RSs$

Example: *with* 3 *concepts and an exhaustive training set, MNLogic has* 6 *RSs if* K *is a conjunction and* 24 *if* K *is a XOR. This grows* ***exponentially*** *with the* # *of concepts!*

.

② **Model-level**: `rsbench` tasks induce RSs in all models!

Table 1. (L) DPL and (R) NN concept confusion matrix on MNAdd-EvenOdd



Quantitatively: Concept F1, accuracy and *collapse*

REFERENCES

[1] Marconato *et al.*, Analysis and Mitigation of RSs, NeurIPS (2023)
[2] Manhaeve *et al.*, DeepProblog, NeurIPS (2018)
[3] Pang Wei Koh *et al.*, Concept bottleneck models, ICML (2020)
[4] Alec Radford *et al.*, CLIP, ICML (2021)

https://unitn-sml.github.io/rsbench/

samuelebortolotti samuele.bortolotti@unitn.it