On the Paradox of Learning to Reason from Data

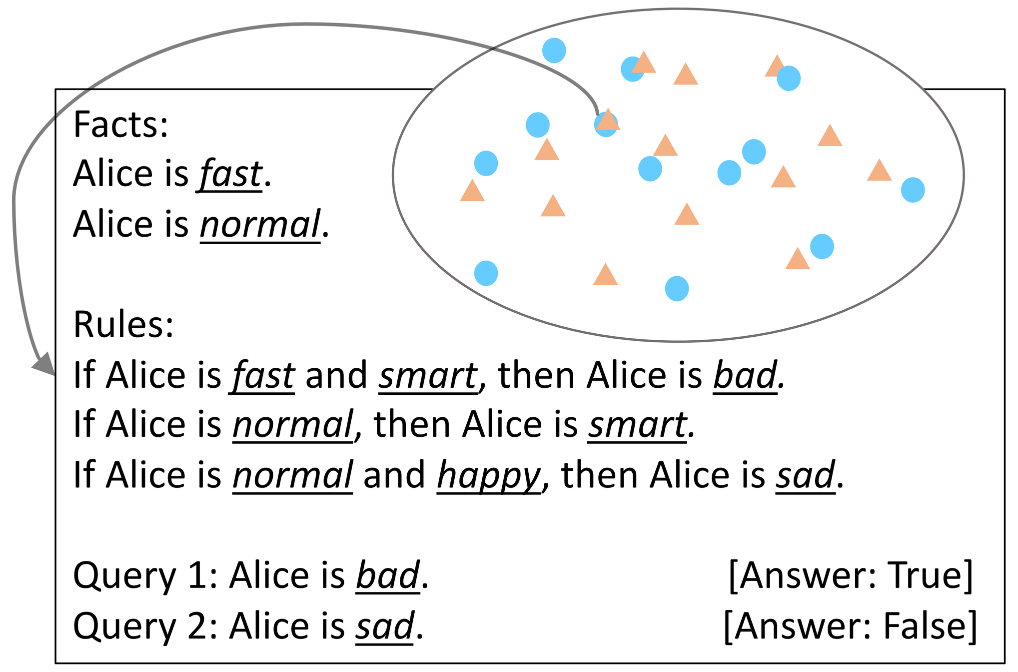
Honghua Zhang Liunian Harold Li Tao Meng Kai-Wei Chang Guy Van den Broeck
University of California, Los Angeles

Can BERT Learn Logical Reasoning?

What is Logical Reasoning

1. *Deductive Reasoning*: the ability to draw conclusions only based on given facts and rules.
2. We say a model can reason if it can reliably emulate a reasoning function (e.g., forward chaining).

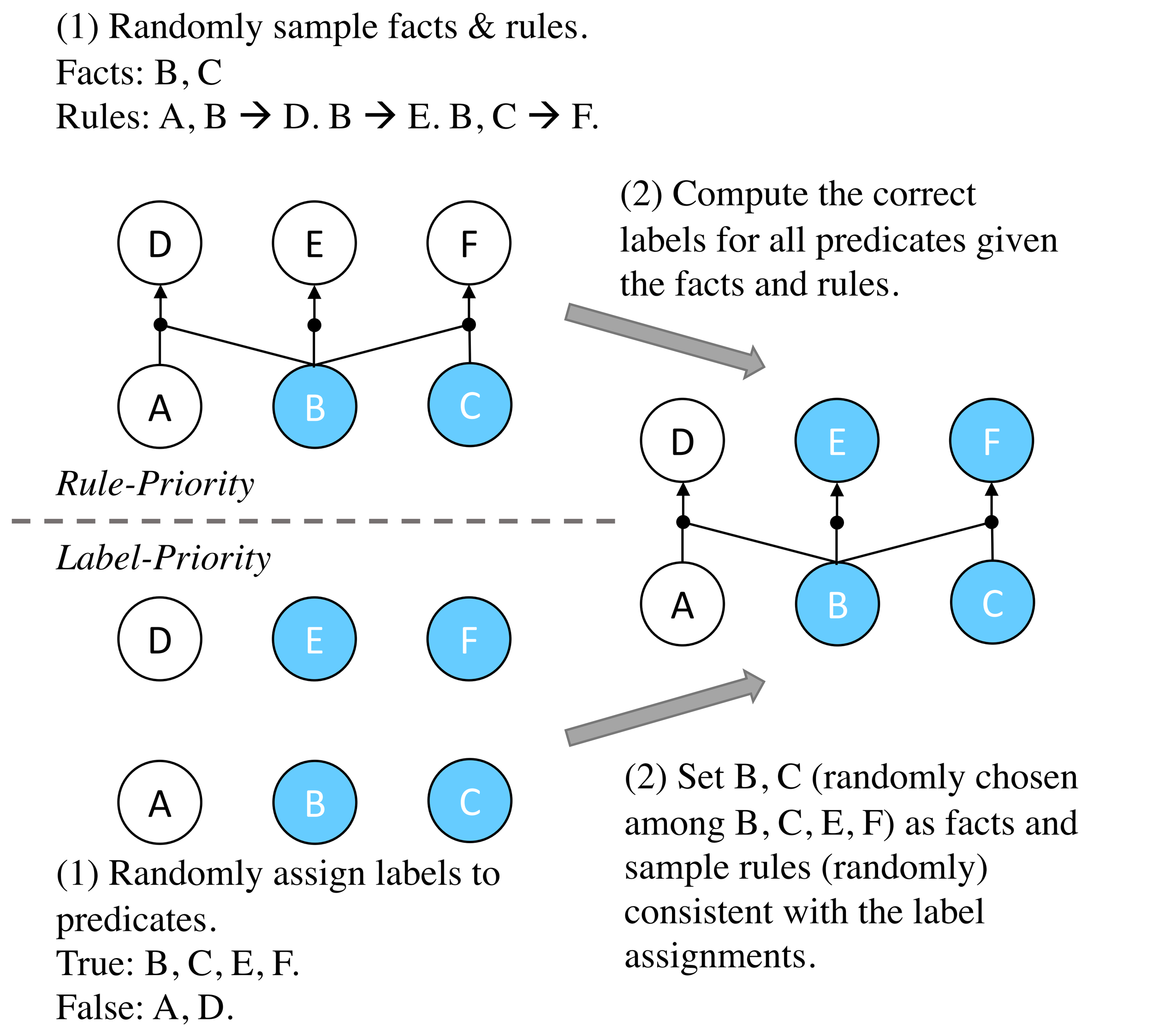
SimpleLogic



1. SimpleLogic is a tractable fragment of logical reasoning problems in propositional logic:
   a. bounded vocabulary (≤ 150) & bounded number of rules/facts (≤ 120).
   b. bounded reasoning steps (≤ 6).
   c. finite domain (≈ 10^{360} examples).
   d. only definite clauses.
   e. predicates are purely symbolic.
2. No language variance: templated language.
3. Examples are self-contained and require no prior knowledge.
4. Transformers *can* solve SimpleLogic:

Theorem. for transformer encoders with n layers and 12 attention heads, there exists a set of parameters that it correctly solves all reasoning problems in SimpleLogic with depth $\leq n-2$.

Sampling Data from SimpleLogic



We construct two datasets RP and LP, each with 280k examples, sampled from Rule-Priority and Label-Priority.

Paradox

Train	Test	0	1	2	3	4	5	6
RP	RP	99.9	99.8	99.7	99.3	98.3	97.5	95.5
	LP	99.8	99.8	99.3	96.0	90.4	75.0	57.3
LP	RP	97.3	66.9	53.0	54.2	59.5	65.6	69.2
	LP	100.0	100.0	99.9	99.9	99.7	99.7	99.0

Test accuracy on LP/RP for the BERT model trained on LP/RP; the accuracy is shown for examples with reasoning depth from 0 to 6. BERT trained on RP achieves almost perfect test accuracy; however, the accuracy drops significantly when it's tested on LP (vice versa).

1. If BERT has learned to reason, it should not exhibit such generalization failure.
2. If BERT has not learned to reason, it is baffling how it achieves near-perfect in-distribution test accuracy.

BERT Learns Statistical Features

What is Statistical Feature

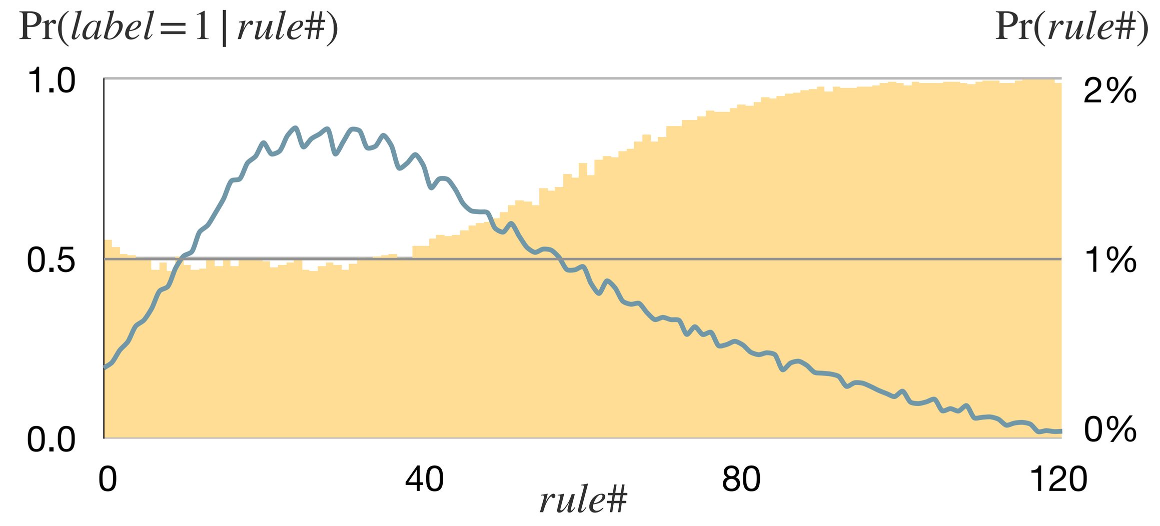
If a certain statistic of examples has a strong correlation with their labels but cannot be used to fully determine the labels, we call it a *statistical feature*.

Statistical Features are Inherent

Monotonicity of entailment: any facts and rules can be freely added to the hypothesis of any proven fact.

The more rules given, the more likely a predicate is proved.

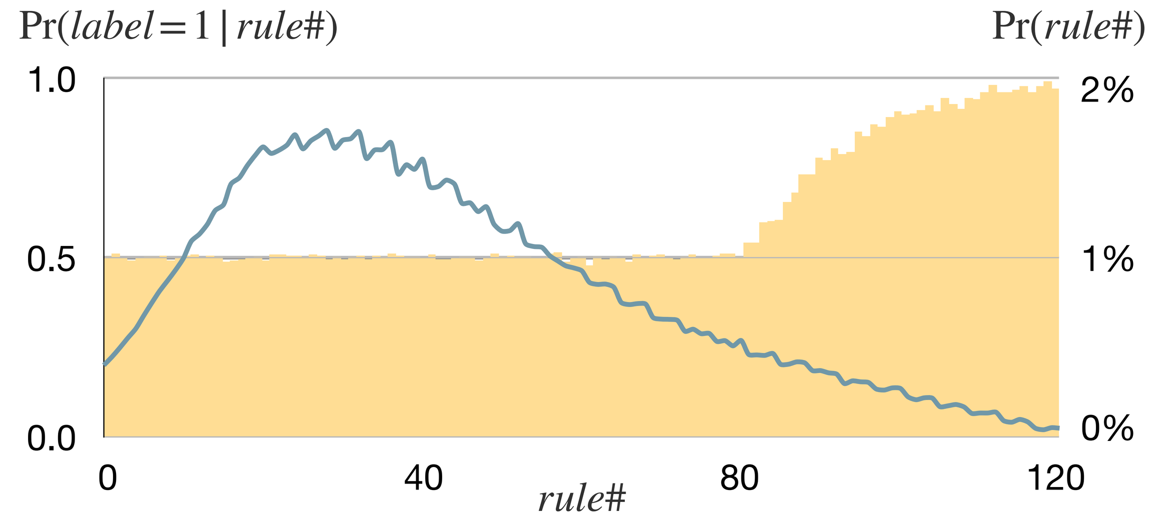
Pr(label = True | rule# = x) should increase (roughly) monotonically with x



Removing Statistical Feature (is Hard)

We down-sample from RP to obtain RP_b such that:

1. Pr(label = True | rule# = x) = 0.5 for all x
2. Pr(rule# = x) stays the same as RP



We need to sample roughly 10x RP before down-sample, taking more than a day on a 40-core CPU. Cost of sampling grows exponentially for jointly removing statistical features.

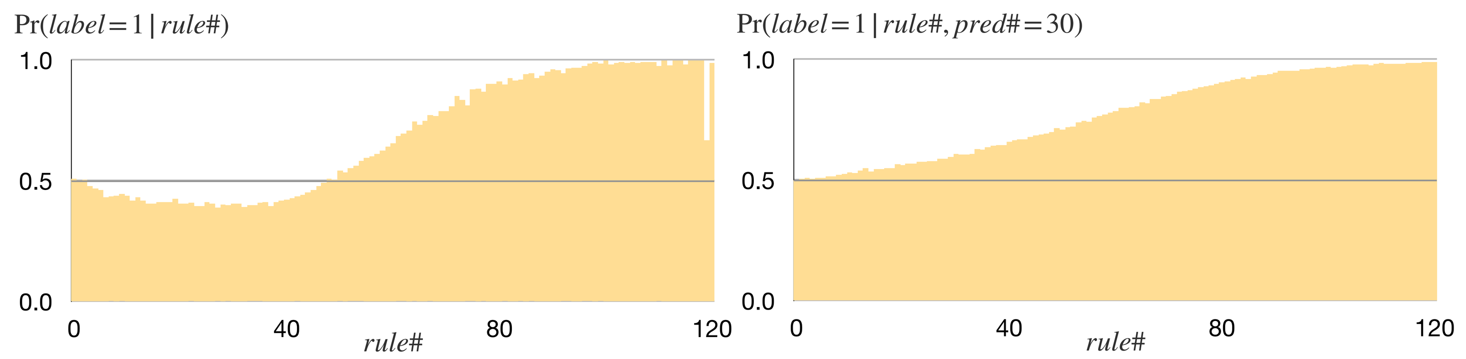
BERT uses Statistical Features

Train	Test	0	1	2	3	4	5	6
RP_b	RP	99.8	99.7	99.7	99.4	98.5	98.1	97.0
	RP_b	99.4	99.6	99.2	98.7	97.8	96.1	94.4
	LP	99.6	99.6	99.6	97.6	93.1	81.3	68.1
RP	RP	99.9	99.8	99.7	99.3	98.3	97.5	95.5
	RP_b	99.0	99.3	98.5	97.5	96.7	93.5	88.3
	LP	99.8	99.8	99.3	96.0	90.4	75.0	57.3

Test accuracy for the BERT model trained on RP/RP_b

1. BERT trained on RP fails to generalize to RP_b, suggesting that BERT leverages rule# to make predictions.
2. BERT trained on RP_b generalizes slightly better, indicating that statistical features inhibit model generalization.

Statistical Features Explain the Paradox



Pr(label = True | rule#) for LP (left) and uniform distributions (right).

Though statistical features are strong signals for in-distribution examples, they vary as the distribution changes.

Main Message

1. We **do not** claim/believe that language models cannot be used to solve any reasoning problems in general :)
2. There is a **fundamental difference** between learning to reason and learning to achieve high performance on NLP benchmarks using statistical features.
3. **Caution** should be taken when we seek to train neural models end-to-end to solve logical reasoning tasks.

All arguments extend to other LMs: e.g., we show that all experiment results hold for T-5.

References: [1] Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *IJCAI*. ijcai.org.
[2] Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentangle- ment in the winograd schema. *arXiv preprint arXiv:2104.08161*.
[3] Roni Khardon and Dan Roth. 1997. Learning to reason. *Journal of the ACM (JACM)*, 44(5):697–725.

Acknowledgements: This work is partially supported by a DARPA PTG grant, NSF grants #IIS-1943641, #IIS-1956441, #CCF-1837129, Samsung, CISCO, and a Sloan Fellowship. This work is supported in part by Amazon scholarship.