How Ambiguous Are the Rationales for Natural Language Reasoning?

A Simple Approach to Handling Rationale Uncertainty

UNIVERSITY OF OXFORD

Hazel Kim
hazel.kim@cs.ox.ac.uk

CLASSTING

Motivation:

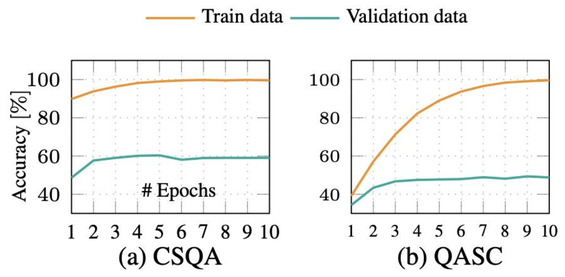
- The quality of rationales is essential in the reasoning capabilities of language models.
- However, obtaining impeccable rationales is often impossible and learning the patterns of rationales is almost improbable.

Contribution:

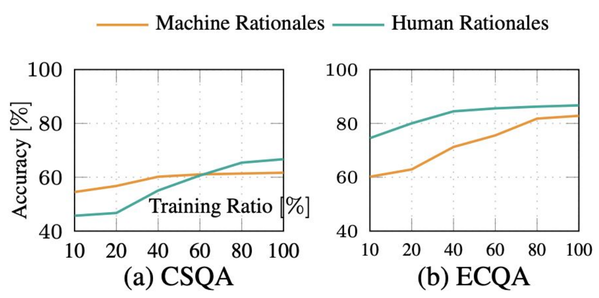
1. Investigate how ambiguous rationales *play in model performances,* both with human-annotated and machine-generated rationales.
2. Assess the *ambiguity of rationales* through the lens of entropy and uncertainty in model prior beliefs,
3. Propose a suitable *two-system reasoning mechanism (**AURA**)* depending on the level of ambiguity.

Observation#1) Ambiguous Rationales and Model Performances

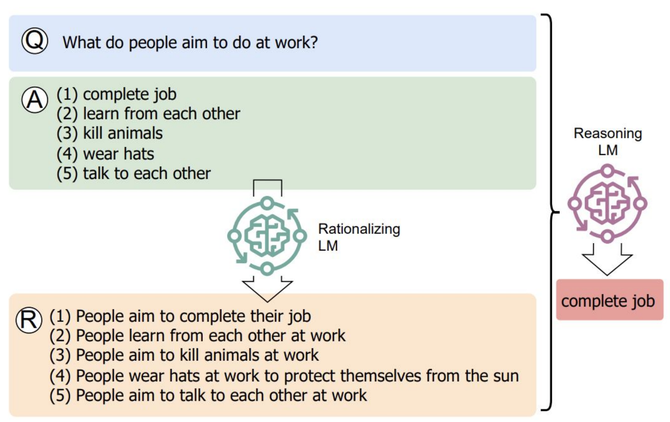
The huge performance mismatch between training and validation sets by epochs. By nature, rationales carry varying normative concepts. This leads models to face the aleatoric uncertainty (i.e., data uncertainty), causing the difficulty of learning the rationales.



Performance of human or machine rationales by training ratios. Though we add up the data points, the performance of the reasoning model remains consistent once it meets certain performance thresholds.



Method:



1. Rationale Entropy

$$H = -\sum_{x} p(x) \log p(x)$$

We use the prior beliefs as informativeness and the posterior beliefs as the probability of the event occurring.

2. Rationale Ambiguity

$$R(x) = \begin{cases} r_{\text{unambiguous}}, & \text{if } h(x_i) < \tau \\ r_{\text{ambiguous}}, & \text{if } h(x_i) \geq \tau \end{cases}$$

We use the expected value of entropies with the entire data points as a threshold to determine ambiguous and unambiguous rationales within the distribution of how data points interact with model training.

3. AURA: Two-System Reasoning

$$\{P(y|x^*, \theta^{(t)}\}_{t=1}^{T=2} \rightarrow \{P(y|W^{(t)})\}_{t=1}^{T=2},$$
$$W^{(t)} \sim P(W|x^t, D)$$

We first use a pre-trained model to train a reasoning module on the entire dataset (Reasoning 1).

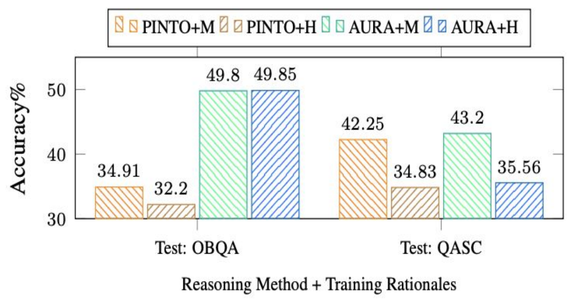
We then train the same pre-trained model again on the selected ambiguous rationales to learn unlearned or under-learned data points that have high entropy scores (reasoning 2). It works as ensemble learning.

Observation #2) Does the quality rationales contribute more than the good reasoning mechanisms to the overall performance?

Robust reasoning method matters more than obtaining quality rationales for predicting right answers. This shows that better reasoning is capable of covering unsatisfactory rationalization.

Performance comparison between human and machine rationales in OoD settings.

For the competitive comparison, we used PINTO, the state-of-the-art method, to compare with AURA as the reasoning method in this experiment.



Observation #3) Do quality rationales provide a better influence on training or inference?

The answer is inference ability. in realistic scenarios, however, we cannot tell if the seemingly ideal rationales are still satisfactory. Hence, AURA with two-system training indeed is strong on both OoD settings and unsatisfactory rationales.

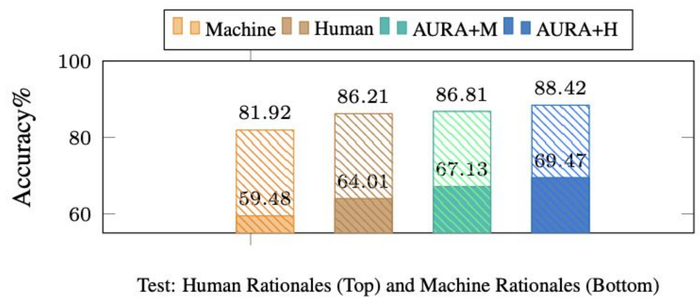
Performance gains by AURA depending on training and testing rationales

M: machine rationales

H: human-written rationales

The legend shows types of training rationales with or without AURA.

Without AURA is standard training.



Observation #4) Robustness of AURA in low-resource and OoD Settings

We test AURA on different training ratios.

We observe AURA outperforms PINTO, the state-of-the-art rationalize-the reason method, in both ID and OoD settings.

Due to the nature of aleatoric uncertainty in rationales, the performance gains by AURA are pretty consistent over different training ratios.