No.2

Transformers know more than they can tell - Investigating the Collatz sequence.

François Charton

Meta

It is generally understood that transformers struggle to learn arithmetic functions (even integer multiplication), models learn shortcuts, fail to generalize. I investigate a complex arithmetic function, predicting distant terms in the Collatz sequence, and show that transformers can learn it to very high accuracy, but incrementally solving the problem for calsses of inputs characterized by their binary representation. This learning pattern is independent of the base used for tokenization. An analysis of model errors unveils a hierarchy of error cases, suggesting that, in latent space, all models are very close to learning the Collatz sequence, no matter the base.